DPTO. DE TEORÍA DE LA SEÑAL Y COMUNICACIONES
UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Nuevos Criterios de Ayuda para Conjuntos de Decisores Cooperativos

Autor: **Vanessa Gómez Verdejo**
Directores: **Dr. Aníbal R. Figueiras Vidal**
Dr. Jerónimo Arenas García

LEGANÉS, 2007

Tesis Doctoral:
Nuevos Criterios de Ayuda para Conjuntos de Decisores Cooperativos

Autor:
Vanessa Gómez Verdejo

Directores:
Dr. Aníbal R. Figueiras Vidal
Dr. Jerónimo Arenas García

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los doctores

Presidente:

Vocales:

Secretario:

acuerda otorgarle la calificación de

Leganés, a

RESUMEN

Aunque en muchas aplicaciones las Redes Neuronales (RRNN) son una herramienta poderosa, en otros problemas (complejos) una única red resulta insuficiente. Para solventar esta dificultad, se puede considerar la combinación de diferentes redes (simples) de modo que se forme un conjunto de RRNN capaz de resolver mejor el problema en cuestión, proporcionando, además, un diseño más sencillo y más fácilmente comprensible, lo que ha ocasionado que su empleo sea cada vez más frecuente. Entre los conjuntos de RRNN destacan, por sus sencillos principios conceptuales y sus contrastadas buenas prestaciones, los métodos de "Boosting", y, especialmente, el algoritmo "AdaBoost".

En esta Tesis Doctoral se partirá del algoritmo "Real AdaBoost" (RA), cuya función de énfasis puede descomponerse en el producto de dos términos, uno relacionado con el error cuadrático de las muestras y otro asociado con la proximidad de las mismas a la frontera. Esta descomposición permite generalizar la estructura de la función de énfasis del RA, introduciendo un parámetro de mezcla ajustable, λ, para controlar el compromiso entre los dos términos de énfasis; el empleo de esta nueva función de énfasis da lugar, como primera aportación, a un nuevo algoritmo que se denomina RA con énfasis ponderado (RA-we, "RA with weigthed emphasis"). Experimentalmente se ha comprobado que si el parámetro de mezcla se selecciona adecuadamente pueden conseguirse mejoras significativas sobre las prestaciones del RA. Sin embargo, no siempre es sencillo encontrar el valor óptimo de λ, y una selección mediante un procedimiento de Validación Cruzada está lejos de aprovechar todo el potencial que el énfasis mixto puede aportar.

Siguiendo esta línea de trabajo, en esta Tesis se exploran, además, dos alternativas para escoger el parámetro de mezcla. La primera de ellas, en lugar de intentar encontrar el mejor valor de λ, combina las salidas de una serie de conjuntos RA-we entrenados con diferentes valores de λ; de este modo, aprovecha la diversidad introducida por el parámetro de mezcla para construir comités de conjuntos RA-we. La segunda de las alternativas propuestas considera una versión generalizada del parámetro de separación del clasificador usado por el algoritmo RA (una correlación ponderada entre las salidas del clasificador y las correspondientes etiquetas), y propone ajustar dinámicamente el parámetro de mezcla

durante el crecimiento del conjunto. Para ello, en cada iteración se selecciona el valor de λ que proporciona un mayor parámetro generalizado de separación.

La idoneidad de estas propuestas es corroborada sobre un conjunto de problemas de decisión binaria, mostrando la efectividad del énfasis mixto, así como de los dos esquemas de selección de λ: comités de conjuntos RA-we y selección dinámica de λ. Además, la comparación de ambas propuestas con esquemas RA clásicos demuestra el interés de los nuevos algoritmos en el ámbito de la construcción de sistemas de múltiples redes.

ABSTRACT

Although Neural Networks (NNs) are an effective tool in many applications, a NN may be inefficient for solving (complex) tasks. To tackle this problem, we may combine a set of NNs in order to construct NN ensemble capable of solving the initial problem, providing an easier design solution and helping to interpret more clearly the resulting machine. The above reasons have increased the interest in this research area during recent years. Among NN ensembles, boosting methods, and in particular AdaBoost, are attractive because of their simple conceptual principles and their good generalization performance.

In this Ph.D. Thesis, we start from the Real AdaBoost (RA) algorithm, where the emphasis function can be decomposed into the product of two factors. The first depending on the quadratic error of each sample, and the second being a function of the "proximity" of the sample to the classification border. This decomposition makes it possible to generalize the structure of the RA emphasis function by introducing an adjustable mixing parameter λ to control the trade-off between both emphasis terms; the algorithm resulting from this proposal is referred to as RA with weighted emphasis (RA-we). Experiments show that a significant improvement over the classical RA performance can be achieved if mixing parameter λ is adequately selected. However, finding the optimal λ is not always an easy task, and using Cross Validation selection methods does not exploit fully the potential that the mixed emphasis function can provide.

Following this research line, this Dissertation also explores two alternatives for selecting the mixing parameter. Rather than trying to find the best value for λ, the first proposal combines the outputs of a number of RA-we networks trained with different values of λ; in this way, we take advantage of the diversity introduced by the mixing coefficient to build committees of RA-we networks. The second approach considers a generalized version of the learner edge defined by the RA algorithm (a weighted correlation between the learners output and the true labels) as an indication of the learner quality, and it proposes to dynamically adjust the mixing parameter during the ensemble growth. In order to do this, we iteratively select the value that provides the learner with the largest generalized edge.

The effectiveness of these two approaches is corroborated over several benchmark bi-

nary decision problems, showing the efficacy of the mixed emphasis approach, as well as the appropriateness of both schemes for selecting λ: (1) committees of RA-we networks, and (2) dynamic λ selection. Finally, we conclude that the algorithms described in this Thesis in comparison to traditional RA algorithms present interesting possibilities for building multi-net systems.

A todos los que, de una manera u otra,
siempre habéis estado a mi lado.

Índice general

Índice de figuras XIV

Índice de cuadros XVI

1. Introducción a los Conjuntos de Redes Neuronales **1**

1.1. Decisión (clasificación) y aprendizaje máquina 2

1.2. ¿Qué es una red neuronal? 3

1.3. Conjuntos de Redes Neuronales 5

1.3.1. Comités de Redes Neuronales 7

1.3.2. Consorcios de Redes Neuronales 9

1.4. "Boosting" 11

1.5. Motivación de esta Tesis Doctoral 13

2. El algoritmo "Real AdaBoost" **17**

2.1. Descripción del algoritmo RA 18

2.2. Propiedades del "Real AdaBoost" 21

2.2.1. Convergencia a cero del error de entrenamiento 22

2.2.2. Análisis del error de generalización 23

2.3. Conclusiones 26

3. "Real AdaBoost" con énfasis mixto **27**

3.1. Análisis de la función de énfasis del RA 28

3.2. Función de énfasis mixto 30
3.3. Construcción de conjuntos de RRNN con énfasis mixto 32
3.4. Análisis de las propiedades del RA-we 35
3.4.1. Convergencia del error de entrenamiento 35
3.4.2. Análisis del error de generalización 35
3.5. Influencia del énfasis mixto en el entrenamiento de los clasificadores base 37
3.6. Selección de λ por validación cruzada 40
3.7. Conclusiones 41

4. Comités de conjuntos RA-we **43**
4.1. Construcción de comités de RA-we 44
4.1.1. Combinación de las salidas de conjuntos RA-we 44
4.1.2. Selección de conjuntos RA-we 51
4.2. Clasificación acelerada de comités de conjuntos RA-we 53
4.2.1. Compatibilidad de los comités con el método de clasificación rápida 54
4.2.2. Procedimiento de clasificación acelerada 57
4.2.3. Reordenamiento de las redes para máximo ahorro computacional 58
4.3. Conclusiones 60

5. Ajuste dinámico de la función de énfasis **63**
5.1. El algoritmo DW-RA 64
5.2. Ventajas de DW-RA frente a RA-we 67
5.2.1. Aceleración de la reducción del error de entrenamiento 67
5.2.2. Mejora en la capacidad de generalización 70
5.3. Conclusiones 72

6. Evaluación de las distintas propuestas **73**
6.1. Descripción de los experimentos 74
6.1.1. Bases de datos empleadas 74

6.1.2. Entrenamiento de los conjuntos RA, RA-we y DW-RA 75
6.1.3. Diferencia estadística entre los resultados: T-test 77
6.2. Selección por CV del parámetro de mezcla 79
6.3. Comités de conjuntos RA-we 83
6.3.1. Prestaciones de los comités lineales 85
6.3.2. Prestaciones de los comités con activación “tanh” 87
6.3.3. Prestaciones de los comités con voto generalizado 88
6.3.4. Prestaciones de los comités con criterio RA-we 90
6.3.5. Evaluación conjunta de los diferentes comités 92
6.4. Clasificación acelerada de comités de conjuntos RA-we 94
6.4.1. Clasificación acelerada con comités lineales 95
6.4.2. Clasificación acelerada con los comités con “tanh” 97
6.4.3. Clasificación acelerada de los comités con criterio RA-we 99
6.5. Prestaciones del algoritmo DW-RA 102
6.5.1. Una primera evaluación 102
6.5.2. Análisis de la velocidad de convergencia 105
6.5.3. Análisis de la capacidad de generalización 107
6.5.4. Aspectos adicionales 109
6.6. Conclusiones: prestaciones de los algoritmos con énfasis mixto 112

7. Conclusiones y Líneas Futuras **115**
7.1. Conclusiones 115
7.2. Líneas de investigación futura 118

A. Análisis del RA-we **121**
A.1. Selección de los pesos de salida 121
A.2. Convergencia del error de entrenamiento 123
A.3. Análisis del riesgo marginal 124
A.4. Análisis de la función de coste de los clasificadores base 126

B. Bases de datos **129**

Índice de figuras

1.1. Esquemas típicos que presentan los distintos tipos de conjuntos: comités y consorcios (sufiguras (a) y (b), respectivamente). 7

1.2. Esquema de un consorcio tipo MoE. 10

2.1. Esquema de un clasificador RA. 19

2.2. Evolución del término t-ésimo de la cota sobre el riesgo marginal en función del parámetro de separación, γ_t, para distintos valores de θ. 26

3.1. Influencia de los distintos tipos de énfasis sobre cada tipo de datos. 30

5.1. Evolución de $h(\delta_t)$ en función del parámetro generalizado de separación. . 71

6.1. Comportamiento $\overline{E}_{clas}$ en función del valor de λ en el problema *Abalone*. 82

6.2. Evolución de $\overline{E}_{clas}$ y $\overline{T}$ en función de β en los comités lineales para los problemas *Image* y *Spam*. 97

6.3. Evolución de $\overline{E}_{clas}$ en función del número de clasificadores evaluados por el método de clasificación acelerada en los comités lineales y el problema *Kwok*. 97

6.4. Evolución de $\overline{E}_{clas}$ en función del número de clasificadores evaluados por el método de clasificación acelerada en los comités criterio RA-we y el problema *Tictactoe*. 101

6.5. Convergencia de $\overline{E}_{clas}$ en los algoritmos RA-se, CV RA-we y DW-RA, en los ocho problemas considerados, cuando cada algoritmo selecciona independiente mediante CV el valor de M empleado. 104
6.6. Convergencia de $\overline{B}_t$ en los algoritmos RA-se, CV RA-we y DW-RA, en los ocho problemas considerados, cuando el valor de M es el seleccionado por el RA-se. 106
6.7. Comportamiento de $\overline{R}_T^{\text{margin}}$ $(\%)$ en función de θ para los algoritmos RA-se, CV RA-we y DW-RA en el problema *Ripley* y fijando M a 48. 109
6.8. Análisis de $\overline{R}_T^{\text{margin}}(\theta)$ en el problema *Image* para $M = 11$; (a) Comportamiento de $\overline{R}_T^{\text{margin}}(\theta)$ en las cercanías de 0; (b) Evolución de $\overline{\rho}_{\min}$ con el número de rondas. 109

Índice de cuadros

2.1. Pseudocódigo de funcionamiento del algoritmo RA. 22

3.1. Pseudocódigo de funcionamiento del algoritmo RA-we. 34

4.1. Pseudocódigo del procedimiento de clasificación acelerada en comités. . . 59

5.1. Pseudocódigo de funcionamiento del algoritmo DW-RA. 66

6.1. Características de las bases de datos empleadas en la evaluación de las técnicas propuestas. 75

6.2. Valores límite del parámetro t del T-test para distintos niveles de certeza. . 79

6.3. Prestaciones presentadas por los algoritmos RA-se y CV RA-we y por la aproximación "omnisciente". 81

6.4. Valores de M seleccionados por cada conjunto RA-we en función del valor de λ empleado por cada uno. 84

6.5. Umbrales empleados por el método de selección de redes en cada problema y en cada tipo de comité. 85

6.6. Prestaciones de los comités lineales (en su versión básica y realizando selección de redes) frente al RA-se. 86

6.7. Prestaciones de los comités con activación "tanh" (en su versión básica y realizando selección de redes) frente al RA-se. 88

6.8. Prestaciones de los comités con voto generalizado (en su versión básica y realizando selección de redes) frente al RA-se. 89

6.9. Prestaciones de los comités con el criterio del RA-we (en su versión básica y realizando selección de redes) frente al RA-se. 91
6.10. Evaluación del mejor de los comités frente al algoritmo RA-se y frente a la aproximación "omnisciente". 93
6.11. Prestaciones del método de clasificación acelerada en los comités lineales. 96
6.12. Prestaciones del método de clasificación acelerada en los comités con "tanh". 98
6.13. Prestaciones del método de clasificación acelerada en los comités con criterio RA-we. 100
6.14. Prestaciones de los algoritmos RA-se, CV RA-we y DW-RA cuando cada algoritmo selecciona mediante un proceso CV independiente el valor de M empleado. 103
6.15. Prestaciones de los algoritmos RA-se, CV RA-we y DW-RA cuando el valor de M se ha fijado al seleccionado por el RA-se. 108
6.16. Errores de clasificación de los algoritmos RA-se, CV RA-we y DW-RA cuando emplean el valor óptimo de M y además, en el caso del algoritmo CV RA-we, el valor óptimo del λ. 110
6.17. Comités de conjuntos RA-we versus selección dinámica. 112

Capítulo 1

Introducción a los Conjuntos de Redes Neuronales

En este primer capítulo de la Tesis Doctoral se presentarán algunos conceptos fundamentales del Aprendizaje Máquina, prestando especial atención a las técnicas de construcción de conjuntos, y más concretamente a las técnicas de "Boosting". Para ello, se revisarán distintas aproximaciones que permiten la resolución del problema de decisión (clasificación), centrándose en las aproximaciones máquina, y más concretamente, en las Redes Neuronales (RRNN); a continuación, se indicarán brevemente las características más significativas de las RRNN, para pasar a describir los conjuntos de RRNN y analizar las ventajas que éstos pueden aportar. Llegados a este punto, se distinguirá entre dos tipos de conjuntos: comités y consorcios; prestando especial atención a los segundos, y, en particular, a las técnicas de "Boosting" y al algoritmo "AdaBoost", punto de partida de esta Tesis Doctoral.

1.1. DECISIÓN (CLASIFICACIÓN) Y APRENDIZAJE MÁQUINA

El problema de decisión plantea cómo elegir entre un cierto número de hipótesis (normalmente exhaustivas y excluyentes) a la vista de un dato, muestra o instancia relacionado con ellas; si las hipótesis son la pertenencia a clases, el problema, llamado de clasificación, se soluciona estableciendo una función capaz de determinar (predecir) la etiqueta (clase) de las instancias no conocidas.

La resolución de un problema de decisión puede abordarse bajo dos perspectivas: la analítica, basada en la teoría estadística, y la que opta por aproximaciones máquina. A continuación, y empleando como guías principales [Bishop, 1995, Duda et al., 2001, Fukunaga, 1990, Haykin, 1999], se presentarán estas dos perspectivas.

La primera de ellas, la perspectiva analítica (véase también [Van Trees, 1968]), conduce a las teorías Bayesiana y frecuentista de la decisión, que, partiendo de la información estadística del problema y una adecuada política de costes, permite el diseño directo de los correspondientes decisores.

En muchos casos no se dispone de dicha información estadística, y una opción es recurrir a su estimación a partir de las muestras disponibles, dando lugar a los diseños semianalíticos. Estos diseños se clasifican en tres tipos, según el método empleado para la estimación de la información estadística necesaria:

1. Paramétricos: consideran una forma analítica para las funciones de densidad, y estiman sus parámetros a partir de los datos disponibles. Su empleo debe hacerse con precaución, ya que una suposición incorrecta del modelo da lugar a clasificadores de bajas prestaciones.

2. No paramétricos: aplican un modelo general capaz de aproximar cualquier tipo de función de densidad, ajustando sus parámetros de acuerdo a las observaciones disponibles; su principal inconveniente es que el modelo implica alto uso de memoria y altas demandas de cómputo en su aplicación. Entre estas técnicas destaca el

método de los k vecinos más próximos, k-NN ("k-Nearest Neighbours"), o el empleo de modelos basados en ventanas de Parzen.

3. Semiparamétricos: se encuentran a medio camino entre los dos diseños anteriores, ya que emplean modelos bastante flexibles para las funciones de distribución, de modo que se consiguen buenas aproximaciones utilizando, al mismo tiempo, pocos parámetros, reduciéndose así el coste computacional. Un modelo difundido es el de mezcla de gaussianas, aplicándose generalmente el algoritmo de Esperanza-Maximización (EM) para la estimación de sus parámetros.

Las aproximaciones máquina permiten diseñar un decisor a partir de un conjunto de muestras, datos o instancias etiquetadas y representativas del problema. El decisor, o clasificador, es una función que divide el espacio en el que se encuentran los datos en distintas hiperregiones, cada una de ellas asociadas a una clase. Estas técnicas se han convertido en las últimas décadas en herramientas muy eficaces para la resolución de múltiples problemas, destacando no sólo en el reconocimiento de patrones, sino también en tareas de estimación (predicción de valores numéricos en lugar de etiquetas).

Entre las técnicas de Aprendizaje Máquina se encuentran las Redes Neuronales (RRNN) ([Bishop, 1995, Duda et al., 2001, Haykin, 1999] son ejemplos de buenas introducciones a las mismas), los Procesos Gaussianos [Rasmussen y Williams, 1996], las Redes Bayesianas [Buntine, 1994], los Árboles de Decisión [Breiman et al., 1984], los Sistemas Expertos (de Reglas) [Michalski, 1980] y los Métodos Basados en Núcleos ("Kernel Methods") [Scholkopf y Smola, 2002], destacando entre los últimos las Máquinas de Vectores Soporte, SVM ("Support Vector Machines") [Burges, 1998, Vapnik, 1995].

1.2. ¿QUÉ ES UNA RED NEURONAL?

La capacidad que el cerebro humano presenta para aprender a partir de un conjunto de estímulos ha motivado el diseño de sistemas que permitan aproximar artificialmente este comportamiento: las Redes Neuronales Artificiales, comúnmente conocidas como Redes

Neuronales (RRNN).

El diseño de las RRNN, imitando las estructuras neuronales naturales, se realiza interconectando un conjunto de unidades de procesamiento, denominadas *neuronas*, a través de una serie de conexiones sinápticas; normalmente estas neuronas se encuentran dispuestas en capas y formando una arquitectura en paralelo, lo que dota a las RRNN de ciertas características estructurales muy ventajosas y que las diferencian de otros sistemas de procesamiento de información; entre ellas se puede destacar:

- **Rapidez**: dada su arquitectura distribuida, las RRNN son sistemas idóneos para el procesamiento en tiempo real mediante procesadores que reproducen la propia arquitectura de la red. El paralelismo otorga a las RRNN gran rapidez en funcionamiento, cuando se implementan así, aunque no por ello en aprendizaje.

- **Robustez**: el conocimiento adquirido se encuentra distribuido por toda la red, de forma que si se daña un número reducido de nodos, las respuestas no se degradan significativamente. Esto dota a las RRNN de una gran robustez frente a deterioro o fallo.

Las RRNN aprenden a partir de los datos que se les suministran; para ello, el valor de las conexiones sinápticas, típicamente pesos, se ajusta mediante un algoritmo de aprendizaje para que, una vez acabado dicho aprendizaje, la red sea capaz de realizar la tarea deseada. Sin embargo, este proceso, conocido como aprendizaje o entrenamiento, suele ser difícil y debe hacerse con particular cuidado para buscar una buena generalización; es decir, para que la red entrenada sea capaz de producir salidas razonables para muestras que no le han sido presentadas con anterioridad, que es lo que en definitiva importa. Es importante señalar que el conocimiento adquirido por la red responde a la información que se le ha presentado, siendo necesario reajustarla o reentrenarla cuando aparecen nuevas muestras o cuando se producen cambios en los datos presentados.

Todas estas propiedades convierten a las RRNN en poderosas herramientas para tareas de decisión, empleándose con éxito en una amplia gama de aplicaciones; si se clasifican según las disciplinas en las que se emplean, destacan:

- Aplicaciones empresariales: reconocimiento de caracteres escritos o del habla, explotación de bases de datos, síntesis de voz desde texto, control de producción en líneas de proceso, inspección de calidad, etc.

- Aplicaciones financieras: predicción en el mercado financiero, análisis de la evolución de los precios, valoración del riesgo de los créditos, detección de fraude (como, por ejemplo, el uso ilegal de tarjetas de crédito) o de falsificaciones (como, por ejemplo, firmas falsificadas), tasación de propiedades, etc.

- Aplicaciones médicas: diagnóstico y tratamiento a partir de síntomas y/o de datos analíticos, detección y evaluación de fenómenos médicos (como, por ejemplo, tumores en mamografías), estimación del coste del tratamiento, predicción de reacciones adversas a los medicamentos, etc.

- Otras aplicaciones: previsión meteorológica, clasificación botánica, predicción en juegos, etc.

1.3. CONJUNTOS DE REDES NEURONALES

Aunque para muchas aplicaciones las RRNN son una herramienta suficiente, para otros problemas (complejos) una (simple) red resulta insuficiente. Para solventar esta dificultad, y bajo la filosofía de dividir un problema complejo en un conjunto de subproblemas más sencillos o bien combinar adecuadamente diferentes soluciones de un mismo problema, se considera la combinación de diferentes redes (simples) de modo que se forme un sistema de RRNN, al que se va a llamar conjunto, capaz de resolver mejor el problema inicial. Nótese que, aunque las RRNN, en general, son máquinas muy potentes, puede ocurrir que a la hora de resolver un determinado problema el tipo de red elegido (o la estructura de la misma) no sea el más adecuado y, por lo tanto, la red resulte ineficiente para su resolución; por este motivo, en muchas ocasiones es más conveniente dividir el problema en distintas regiones, facilitando así el entrenamiento de la red, o combinar diferentes

soluciones (aunque éstas no sean por sí solas lo suficientemente buenas) para conseguir resolver el problema que inicialmente parecía complejo.

Los principios que siguen los sistemas de RRNN es similar a los que presenta el cerebro humano a la hora de realizar una tarea. Considérese, por ejemplo, el funcionamiento de la visión humana [Ratey, 2002]; cuando los rayos de luz inciden en el ojo, éste transforma la señal luminosa en energía que es enviada al cerebro a través del nervio óptico; una vez que llega a la corteza visual del cerebro, la información es repartida entre diferentes regiones del cerebro que se encargan del análisis de distintos rasgos (algunas extraen bordes o líneas, otras se encargan del color, del tamaño o de la orientación), y es la combinación de los resultados de estos análisis la que permite que se adquiera una información visual completa de lo que se tiene delante.

En general, un conjunto de RRNN no sólo es capaz de ofrecer soluciones considerablemente mejores que cualquiera de las proporcionadas por las redes aisladas, sino que, además, puede implicar una reducción considerable del tiempo de entrenamiento y la obtención de un sistema más fácilmente comprensible y modificable [Sharkey, 1996], lo que ha ocasionado que, desde que se propusieron en la década de los sesenta (véase, por ejemplo [Nilsson, 1965]), su empleo haya sido cada vez más frecuente, sobre todo a partir del trabajo de Hansen y Salomon [Hansen y Salomon, 1990].

Los conjuntos de RRNN pueden clasificarse atendiendo a distintos criterios, como puede comprobarse en [Sharkey, 1999] o en [Haykin, 1999]; aquí se han clasificado atendiendo a la manera que tienen las redes de ayudarse entre sí para resolver el problema. Según este criterio, los conjuntos de RRNN pueden dividirse en comités y consorcios:

- En un **comité**, todas las redes resuelven el mismo problema y, con una adecuada combinación de sus salidas, permiten obtiener soluciones mucho más precisas que cualquiera de las soluciones individuales (véase la Subfigura 1.1(a)).

- En un **consorcio**, las redes componentes cooperan entre sí, siendo la contribución de todas las redes la que proporciona la solución al problema global. Por un lado, existen esquemas en los que, siguiendo el principio de divide y vencerás, el problema

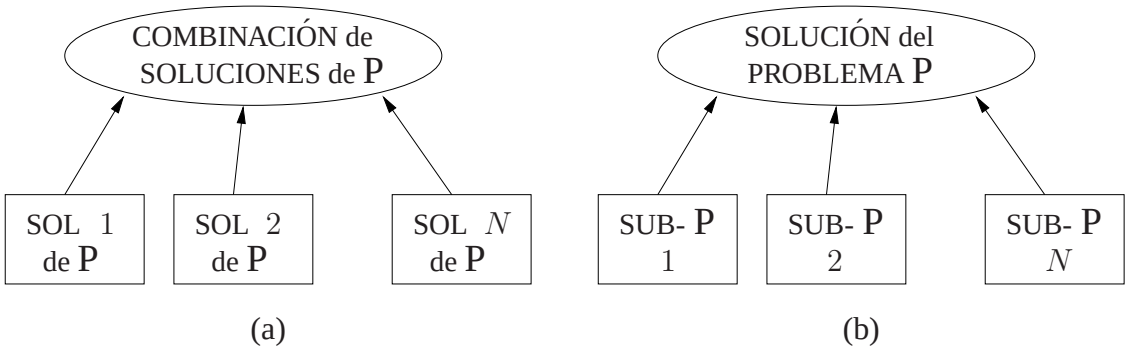

Figura 1.1: Esquemas típicos que presentan los distintos tipos de conjuntos: comités y consorcios (sufiguras (a) y (b), respectivamente).

a resolver se divide en un número de subproblemas, cada uno de ellos asignado a un experto diferente (véase la Subfigura 1.1(b)), como es el caso de los conjuntos modulares o mezclas de expertos, “Mixture of Experts” (MoE) [Jacobs et al., 1991, Jordan y Jacobs, 1994]. Por otro lado, existen métodos, como el “Boosting”, en los que se añaden redes al conjunto de modo que se mejore la solución proporcionada por las previamente incorporadas.

Cabe destacar que, aunque estos esquemas presenten filosofías claramente distintas, no son estructuras exclusivas, ya que podrían combinarse entre sí, dando lugar, por ejemplo, a un consorcio de redes, cada una de las cuales es, a su vez, una combinación modular.

De ahora en adelante, se centrará la atención en los métodos de construcción de conjuntos, analizándolos por separado para cada uno de los tipos de conjuntos que se acaban de presentar.

1.3.1. Comités de Redes Neuronales

Como ya se ha indicado, los comités se basan en la combinación de distintas redes de manera que (idealmente) la red global supere en prestaciones a cualquiera de las redes aisladas. Esta propiedad se puede explicar considerando que cada red componente presenta una serie de limitaciones a la hora de resolver el problema deseado, cometiendo, por ello,

una serie de errores; si se realiza una adecuada combinación de las salidas de las redes, de modo que los errores cometidos por una de ellas sean corregidos por las otras, se puede minimizar el número de errores, obteniendo una solución de mayor precisión.

Por este motivo, para una construcción adecuada de un comité de RRNN, es deseable que los errores que cometen las redes puedan ser compensados por las otras, para lo cual estas redes deben generalizar de forma diferente [Krogh y Vedelsby, 1995, Rosen, 1996], i.e., se precisa una baja correlación entre los errores de las redes componentes. Esta característica, conocida como diversidad, es la que explotan la mayoría de las técnicas de construcción de comités.

Las maneras más sencillas de forzar diversidad entre las redes que van a componer el comité consisten en emplear distintas condiciones iniciales en el entrenamiento, distintas topologías, distintos algoritmos de entrenamiento o, simplemente, mediante técnicas de remuestreo o filtrado que permiten que cada clasificador emplee un conjunto de entrenamiento diferente. Esta última manera de crear diversidad es la que ha dado lugar a la mayoría de métodos de construcción de comités, destacando:

- **Obtención de diferentes conjuntos de entrenamiento**: estos métodos [Sharkey et al., 1996, Sharkey y Sharkey, 1997] generan distintos conjuntos de entrenamiento empleando datos procedentes de distintas fuentes o utilizando distintas técnicas de preprocesado.

- **Utilización de técnicas elaboradas de muestreo**: estas técnicas consisten en que cada una de las redes base es entrenada con un subconjunto del conjunto inicial de datos de entrenamiento. Entre ellas se encuentran los métodos de “Bootstrapping”, destacando el algoritmo de “Bagging” (“Bootstrap AGGregatING”) propuesto por Breiman en 1996 [Breiman, 1996]. En este algoritmo, cada red se entrena con un subconjunto de datos del conjunto original, que es creado mediante un remuestreo con reemplazamiento, y se obtiene la salida del conjunto realizando el promedio de las salidas de todas las redes.

- **Empleo de conjuntos de entrenamiento disjuntos**: estos métodos siguen una idea

similar al "Bootstrapping", ya que cada red se entrena con distintos subconjuntos de datos creados mediante remuestreo, pero, en este caso, se emplea remuestreo sin reemplazamiento, de modo que los subconjuntos creados son disjuntos (véase por ejemplo [Sharkey et al., 1996]). El problema de este tipo de métodos es que cuando no se dispone de un elevado número de datos de entrenamiento, el tamaño de los subconjuntos es bastante reducido y se suelen deteriorar las prestaciones.

- **Métodos de remuestreo adaptativo**: Schapire mostró cómo una serie de redes débiles pueden convertirse en una red fuerte como resultado de entrenar a sus miembros con un conjunto de datos que ha sido filtrado por los miembros que lo formaban previamente, dando origen a los métodos de "Boosting" [Schapire, 1990]. Cabe aclarar que, aunque este tipo de métodos pueden verse como un tipo de comités en los que se varía mediante remuestreo el conjunto de datos de entrenamiento, también pueden considerarse como consorcios ya que las redes se van añadiendo al conjunto iterativamente de modo que mejoren la solución proporcionada por las anteriores redes.

Además del criterio empleado para el diseño de las diversas redes, es necesario seleccionar un criterio para combinar sus salidas, siendo este criterio tanto o más importante que el diseño de las redes. Existen para ello múltiples opciones (según se discute en [Jacobs, 1995, Xu et al., 1992] o, más extensamente, en [Kuncheva, 2004]), siendo frecuentes las combinaciones lineales, ya sea mediante un promediado directo de las salidas de todas las redes o empleando un conjunto de pesos a ajustar.

1.3.2. Consorcios de Redes Neuronales

A diferencia de los comités, en un consorcio de RRNN los componentes del conjunto colaboran entre sí, siendo necesarias las aportaciones de todos para obtener la solución del problema. Dentro de esta clase de conjuntos, normalmente se encuentran dos tipos de esquemas: los conocidos como mezclas de expertos ("Mixture of experts", MoE) y los construidos mediante "Boosting". En este apartado se detallarán las MoE, dejando los

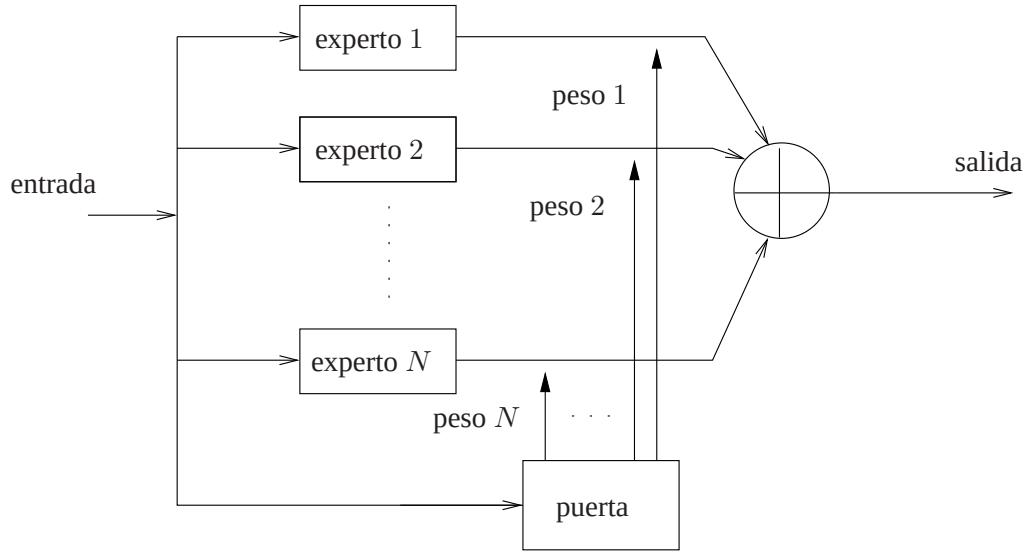

Figura 1.2: Esquema de un consorcio tipo MoE.

métodos de “Boosting” para la siguiente sección.

En una MoE [Jacobs et al., 1991] el espacio en el que se encuentran los datos es divido (de forma dura o blanda) en una serie de subespacios, asignando cada uno de éstos a una red o experto diferente. De este modo, cada experto se entrena para aprender una parte del problema. La salida del conjunto se obtiene con la ayuda de una puerta que asigna un peso a la salida de cada experto en función de la localización del dato de entrada, obteniéndose la salida global del conjunto como combinación lineal ponderada de las salidas de los expertos (véase la Figura 1.2). De este modo, la puerta da más importancia a las salidas de los expertos que se han especializado en la región de espacio en la que se encuentra el dato que se desea clasificar.

Si las MoE dividen el espacio de entrada en regiones, se podría avanzar un paso más dividiendo cada una de estas regiones en subregiones, formando así las llamadas MoE jerárquicas, “Hierarchical Mixture of Experts” (HMoE) [Jordan y Jacobs, 1992]. En este caso, el espacio de entrada se divide en una red de subespacios, cada uno de ellos asignado a un experto, y son necesarias varias puertas, situadas jerárquicamente, para obtener la salida final del conjunto.

1.4. "BOOSTING"

La idea que subyace tras los métodos de "Boosting" tiene sus orígenes en la llamada Teoría PAC ("Probably Approximately Correct") [Valiant, 1984], según la cual un conjunto de clasificadores con una tasa de acierto ligeramente mejor que el propio azar ("débiles") pueden combinarse para dar lugar a un clasificador con unas prestaciones arbitrariamente elevadas ("fuerte").

Siguiendo esta teoría, en 1990 Schapire introdujo el primer método de "Boosting" [Schapire, 1990], y son muchos los algoritmos que se han propuesto desde entonces siguiendo esta idea, destacando entre todos ellos el "AdaBoost" ("ADAptive Boosting"), propuesto por Freund y Shapire en 1996 [Freund y Schapire, 1996]. Dicho algoritmo consiste en entrenar consecutivamente clasificadores débiles empleando poblaciones en las que se enfatizan las muestras de acuerdo a su error, formando el conjunto por combinación lineal ponderada de las salidas de estos clasificadores débiles.

Este algoritmo, también denominado "AdaBoost.M1", se propuso inicialmente para la resolución de problemas de clasificación binarios, pero, a partir de él, han surgido múltiples versiones que permiten su aplicación a otros problemas; entre ellas se encuentran el "AdaBoost.M2", para la resolución de problemas de clasificación multiclase, el "AdaBoost.R", que permite la resolución de problemas de regresión, o el "Real AdaBoost" (RA), empleado también para clasificación binaria, pero que, a diferencia del "AdaBoost", combina clasificadores cuya salida es un número real y no simplemente un indicador de clase. De este modo, el RA tiene en cuenta no sólo el criterio de cada componente, sino también la "confianza" de dichas predicciones.

Aunque el "AdaBoost" es el algoritmo de "Boosting" más conocido, hay muchos otros algoritmos de esta clase que también han cobrado importancia; entre ellos se destacan:

- Algoritmos de "Arcing" ("Adaptative Resample and Combining"): con este nombre se conoce una familia de algoritmos de "Boosting" propuestos por Breiman en 1999 [Breiman, 1999b]. Los fundamentos de esta familia de algoritmos, y concretamente del Arc-x4, son similares a los del "AdaBoost", salvo en dos aspectos: la salida del

conjunto de obtiene mediante combinación lineal no ponderada de todos los clasificadores base (i.e., se asigna igual peso a todas las redes base); y, aunque se enfatizan las muestras erróneas, lo hace de modo diferente al “Boosting” tradicional[1].

- Algoritmos que maximizan el margen de clasificación: intentando evitar los problemas de sobreajuste que el “AdaBoost” puede presentar, se han propuesto implementaciones que maximizan directamente el margen de clasificación y, a la vez, introducen un término de regularización que reduce dicho sobreajuste. La primera de estas propuestas se corresponde con el algoritmo Arc-GV (“Arcing Game Value”) propuesto por Breiman [Breiman, 1999b], a la que le siguen una serie de algoritmos propuestos por Rätsch: ν-Arc [Rätsch et al., 2000], AdaBoost$_{Reg}$ [Rätsch et al., 2001], AdaBoost$_{\rho}$ y AdaBoost* [Rätsch y Warmuth, 2005].

- Algoritmos de aprovechamiento máximo (“leveraging”): bajo este nombre (“aprovechamiento máximo”) se encuentran un conjunto de algoritmos de “Boosting” caracterizados por emplear como función de énfasis la derivada de la función de coste aplicada para seleccionar los pesos de salida de cada clasificador [Meir y Rätsch, 2003]; este diseño permite relacionar estos métodos con técnicas de optimización numérica para poder probar la convergencia de los mismos [Rätsch et al., 2002]. Dentro de este grupo de algoritmos se encuentran el propio “AdaBoost” y algoritmos como el “Least-Square-Boost” [Friedman, 2001] o el “Logic-Boost” [Friedman et al., 2000], propuestos ambos para la resolución de problemas de regresión.

[1]Mientras que el “AdaBoost” emplea como función de énfasis una función de tipo exponencial, los algoritmos Arc-xn se caracterizan por emplear como función de énfasis un polinomio de orden n.

1.5. MOTIVACIÓN DE ESTA TESIS DOCTORAL

La propuesta del "AdaBoost" no sólo ha dado lugar a la creación de nuevos algoritmos que siguen la misma filosofía, sino que también ha motivado que muchos autores analizaran su comportamiento para intentar justificar sus buenas prestaciones. La mayoría de estos estudios se han centrado en su convergencia, en su capacidad de generalización, o en relacionar su funcionamiento con el de las SVM (véase un resumen de estos estudios en [Freund y Schapire, 1999], o, más detalladamente, en [Meir y Rätsch, 2003]).

Por otro lado, y a diferencia de las anteriores líneas, Breiman estudió la influencia que tenía utilizar distintos tipos de funciones de énfasis; concretamente, empleando los algoritmos "Arcing" y comparándolos con el "AdaBoost", llegó a la conclusión de que el éxito del "AdaBoost" se debe únicamente al hecho de enfatizar las muestras erróneas, siendo poco relevante la función de énfasis empleada [Breiman, 1999a].

A pesar de los estudios de Breiman, y aunque el criterio empleado por el "AdaBoost" para enfatizar las muestras conduce a muy buenos resultados, hay cuestiones que siguen abiertas:

- Estudios sobre el "AdaBoost" [Arenas-García et al., 2003] han puesto de manifiesto que, tras varias iteraciones, el énfasis acaba centrándose en aquellas muestras más cercanas a la frontera de decisión. ¿Por qué dedicar, entonces, la atención del algoritmo sobre las muestras erróneas, y no enfatizar directamente las que están próximas a la frontera?. Esta idea ya ha sido utilizada para mejorar las prestaciones de clasificadores tipo SVM mediante selección dinámica de muestras críticas [Lyhiyaoui et al., 1999, Mora-Jimenez et al., 2003].

- Otros estudios [Rätsch et al., 2001] han mostrado que en presencia de "outliers" (muestras fuera del margen), el "AdaBoost" enfatiza reiteradamente, iteración tras iteración, estas muestras hasta que consigue clasificarlas correctamente, proporcionando normalmente soluciones sobreajustadas que, por tanto, presentan mala generalización. Aunque se han propuesto versiones modificadas que plantean soluciones regularizadas para evitar este efecto, como es el caso del AdaBoost$_{Reg}$

> [Rätsch et al., 2001], ¿no podría evitarse este efecto si se enfatizaran los patrones cercanos a la frontera en vez de, o además de, los erróneos?

Estas cuestiones han motivado que nuestra atención se dedicase a la función de énfasis del "AdaBoost", cuyo análisis demuestra que dicha función se encuentra compuesta de dos factores, uno relacionado con el error cuadrático de las muestras y otro asociado con la proximidad de las mismas a la frontera.

Esta primera observación sirvió de inicio al trabajo que se presenta en esta Tesis Doctoral, planteándose si la manera que tiene el "AdaBoost" de combinar estos dos términos es la más adecuada; llegándose después a proponer una versión modificada del "AdaBoost" que incluye una nueva función de énfasis que combina mediante un parámetro de mezcla los factores asociados con el error y la proximidad a la frontera. Se comprobará con ello que el criterio empleado por el "AdaBoost" para enfatizar la población no siempre es el mejor, y que, en la mayoría de los casos, una combinación adecuada de estos términos de énfasis puede aportar reducciones en la tasa de error, mayores velocidades de convergencia, y, además, puede ayudar a paliar el fenómeno de sobreajuste. Este resultado apareja la resolución de un problema de diseño, consistente en establecer un método para la adecuada selección del valor del parámetro de mezcla; lo que constituirá el núcleo principal de esta Tesis Doctoral.

Una primera aproximación para elegir el parámetro de mezcla, basada en un proceso de validación cruzada, proporciona ventajas frente al "AdaBoost", pero no aprovecha al máximo las posibilidades que la nueva función de énfasis puede aportar. Por este motivo, se exploran dos alternativas. La primera de ellas no opta por buscar el mejor valor del parámetro de mezcla, sino que aprovecha la diversidad existente entre conjuntos construidos con distintos valores del parámetro de mezcla y, mediante su adecuada combinación, mostrará como se pueden obtener mejores resultados que los presentados por cualquiera de los conjuntos individualmente. La segunda de estas alternativas, basándose en los fundamentos teóricos que justifican el buen funcionamiento del "AdaBoost", realizará una selección dinámica y automática de parámetro de mezcla, de modo que en cada iteración se elija el valor del parámetro de mezcla que resulte mejor para la progresión de las presta-

ciones del conjunto en construcción.

Además, dado el alto coste computacional que requiere la evaluación de los clasificadores resultantes, especialmente en el caso de la primera de las alternativas, se propondrá un método de clasificación acelerada que permita la reducción del coste computacional durante la fase operacional de la red.

El trabajo que se acaba de describir se presenta en los Capítulos 3, 4 y 5 de esta memoria y constituye, junto con la evaluación experimental de las propuestas (Capítulo 6), el núcleo principal de esta Tesis Doctoral. El contenido de la memoria se complementa, en el Capítulo 2, con una introducción general al algoritmo "Real AdaBoost", punto de partida de la investigación desarrollada, y concluye, en el Capítulo 7, con una reflexión sobre el trabajo realizado y su relevancia, así como sobre las líneas de investigación que siguen abiertas y que sería interesante abordar en un futuro próximo.

CAPÍTULO 2

EL ALGORITMO "REAL ADABOOST"

Desde que Freund y Schapire propusieron el algoritmo "AdaBoost" [Freund y Schapire, 1996], son muchas las versiones que han surgido mejorando y extendiendo sus prestaciones (en [Meir y Rätsch, 2003] se recogen algunas de las más relevantes), y aunque todas ellas emplean los mismos principios que la original, presentan claras diferencias en su implementación.

Entre estas propuestas destaca la versión aportada por Schapire y Singer [Schapire y Singer, 1999] para clasificación binaria. Esta versión, conocida como "Real AdaBoost" (RA), extiende la original al caso en el que los clasificadores que componen el conjunto tienen un rango de salida continuo, y no limitado al conjunto de valores $\{-1, 1\}$; de este modo, el valor de salida de los clasificadores no sólo indica la clase a la que se asigna el dato, sino que además indica la "confianza" de dicha predicción.

Tal y como se señaló al final del capítulo anterior, los algoritmos de "Boosting" que se proponen en esta Tesis Doctoral emplean el algoritmo "AdaBoost", y concretamente la versión correspondiente al RA, como punto de partida. Por este motivo, antes de presentar dichas propuestas, en este capítulo se va a analizar en detalle el algoritmo RA, describiendo, en primer lugar, su funcionamiento, y analizando, a continuación, sus principales características.

2.1. DESCRIPCIÓN DEL ALGORITMO RA

Considérese un problema de clasificación binaria del que se dispone de un conjunto de muestras con sus correspondientes etiquetas

$$S = \{(\mathbf{x}^{(1)}, d^{(1)}), ..., (\mathbf{x}^{(L)}, d^{(L)})\}, \qquad d^{(l)} \in \{-1, +1\}, \quad l = 1, ..., L$$

y el objetivo es construir una función capaz de clasificar nuevas muestras lo más correctamente posible.

Para resolver este problema, el RA propone entrenar una serie de clasificadores, llamados clasificadores base, y combinarlos de tal modo que se obtenga un clasificador con elevadas prestaciones. Para ello, durante una serie de rondas, $t = 1, ..., T$, entrena un clasificador que implementa una función $o_t(\mathbf{x}) \in [-1, 1]$, asignándole un peso de salida, α_t, y lo añade al conjunto de modo que la salida global del sistema, $f_T(\mathbf{x})$, se obtenga como combinación lineal ponderada de todos los clasificadores base (véase la Figura 2.1)

$$f_T(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t o_t(\mathbf{x}) \tag{2.1}$$

De este modo, cuando llegue un nuevo dato la etiqueta asignada por el sistema estará dada por

$$\hat{d}(\mathbf{x}) = \text{sign}\,[f_T(\mathbf{x})] \tag{2.2}$$

Cada uno de los clasificadores base se entrena para minimizar el error cuadrático medio del conjunto de datos original enfatizado, es decir, el t-ésimo clasificador se entrena para que minimice la siguiente función de coste

$$C_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2 \tag{2.3}$$

donde $D_t(\cdot)$ es la función de énfasis que indica la importancia que el clasificador debe asignar a cada muestra. El primer clasificador asigna a todos los datos la misma importancia:

$$D_1(\mathbf{x}^{(l)}) = 1/L, \quad l = 1, ..., L$$

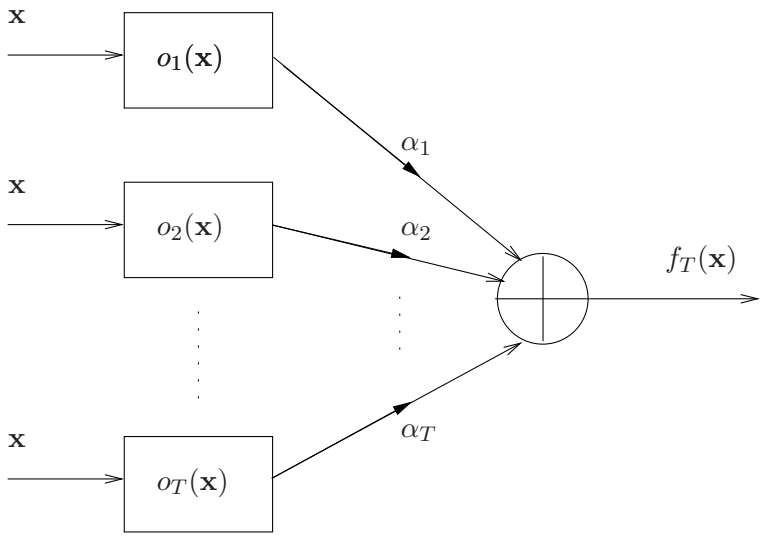

Figura 2.1: Esquema de un clasificador RA.

y en cada ronda esta función se actualiza para que el siguiente clasificador preste mayor atención a los patrones que se han clasificado erróneamente por los clasificadores anteriores; para ello, se emplea la regla de actualización

$$D_{t+1}(\mathbf{x}^{(l)}) = \frac{D_t(\mathbf{x}^{(l)}) \exp[-\alpha_t o_t(\mathbf{x}^{(l)}) d^{(l)}]}{Z_t} \tag{2.4}$$

donde Z_t es un factor de normalización, que se calcula mediante

$$Z_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) \exp[-\alpha_t o_t(\mathbf{x}^{(l)}) d^{(l)}] \tag{2.5}$$

para asegurar que $\sum_{l=1}^{L} D_{t+1}(\mathbf{x}^{(l)}) = 1$.

El cálculo del peso de salida asociado al clasificador t-ésimo, α_t, se realiza de modo que se minimice la siguiente cota del error de entrenamiento, E_t^S, tras la correspondiente ronda:

$$B_t = \frac{1}{L} \sum_{l=1}^{L} \exp[-f_t(\mathbf{x}^{(l)}) d^{(l)}] \geq \frac{1}{2L} \sum_{l=1}^{L} \mid \operatorname{sign}[f_t(\mathbf{x}^{(l)})] - d^{(l)} \mid = E_t^S \tag{2.6}$$

donde f_t es la salida parcial del sistema en la ronda t-ésima, es decir,

$$f_t(\mathbf{x}) = \sum_{t'=1}^{t} \alpha_{t'} o_{t'}(\mathbf{x}) \tag{2.7}$$

Tal y como puede verse en [Schapire y Singer, 1999], el procedimiento seleccionado para la minimización de B_t conduce a distintos métodos para el cálculo de los $\{\alpha_t\}$, dando lugar a distintas implementaciones del algoritmo. Entre este conjunto de posibilidades se ha escogido el método que emplea el "AdaBoost" original, válido para el RA cuando el rango de valores de salida de los clasificadores se encuentra en el intervalo $[-1, 1]$. En este caso, en lugar de minimizar directamente B_t, se recurre a la minimización de la siguiente cota superior[1]:

$$B_t = \frac{1}{L}\sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] \exp[-\alpha_t o_t(\mathbf{x}^{(l)})d^{(l)}] \leq$$

$$\leq \frac{1}{L}\sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]\left[\frac{1+o_t(\mathbf{x}^{(l)})d^{(l)}}{2}\exp(-\alpha_t) + \frac{1-o_t(\mathbf{x}^{(l)})d^{(l)}}{2}\exp(\alpha_t)\right] \quad (2.8)$$

Nótese que cuando el rango de salida de los clasificadores está limitado a los valores $\{-1, 1\}$, es decir, cuando se trata de la versión original del "AdaBoost", se verifica la igualdad en (2.8).

Para minimizar la cota sobre B_t, se deriva la expresión anterior con respecto a α_t y se iguala a 0; de este modo, y tras algunas manipulaciones elementales, puede obtenerse la siguiente expresión cerrada para el cálculo de α_t

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1+\gamma_t}{1-\gamma_t}\right) \quad (2.9)$$

donde γ_t, denominado "edge" o parámetro de separación del t-ésimo clasificador, se calcula como

$$\gamma_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)})o_t(\mathbf{x}^{(l)})d^{(l)} \quad (2.10)$$

[1]La demostración de la existencia de dicha cota se obtiene mostrando que la función $f(x) = \exp[-\alpha_t x] - \frac{1+x}{2}\exp(-\alpha_t) - \frac{1-x}{2}\exp(\alpha_t)$ es positiva en el intervalo $[-1, 1]$, lo que se demuestra viendo que $f(x)$ se anula en $x = \pm 1$ y que su derivada segunda, $f''(x)$, es positiva $\forall x$; y a partir de este resultado, es inmediato que $\exp[-\alpha_t o_t(\mathbf{x}^{(l)})d^{(l)}] \leq \frac{1+o_t(\mathbf{x}^{(l)})d^{(l)}}{2}\exp(-\alpha_t) + \frac{1-o_t(\mathbf{x}^{(l)})d^{(l)}}{2}\exp(\alpha_t)$ para $o_t(\mathbf{x}^{(l)})d^{(l)} \in [-1, 1]$, de donde se deduce fácilmente (2.8).

e indica la calidad de cada uno de los clasificadores, ya que mide, sobre el conjunto de entrenamiento, la correlación existente entre las salidas y las correspondientes etiquetas, enfatizando la población de acuerdo con $D_t(\cdot)$.

Los detalles del desarrollo matemático que conduce a (2.9) y (2.10) se encuentran recogidos en el Anexo A.1; concretamente, dicho desarrollo contempla un caso más general que incluye al RA como caso particular del algoritmo que se propondrá en el capítulo siguiente.

El resto de posibilidades propuestas en [Schapire y Singer, 1999] para el calculo del conjunto de pesos $\{\alpha_t\}$ consisten en minimizar otras cotas de B_t, o bien minimizar directamente B_t mediante métodos numéricos. La elección de (2.9) se ha realizado tras comprobar que el empleo de un criterio u otro no produce diferencias apreciables en el comportamiento global del conjunto, es decir, las prestaciones finales son similares. De esta forma, disponer de una expresión analítica para $\{\alpha_t\}$ simplifica el diseño y minimiza el coste computacional, sin afectar, por ello, a las prestaciones obtenidas.

En el Cuadro 2.1 se recoge el pseudocódigo que describe paso a paso el funcionamiento de este algoritmo.

2.2. PROPIEDADES DEL “REAL ADABOOST”

Antes de tratar el caso particular del RA, cabe decir que son muchos los autores que han analizado el comportamiento del “AdaBoost” para justificar su buen funcionamiento, obteniendo dos explicaciones que justifican las excelentes prestaciones de este algoritmo. La primera de ellas se basa en la plausible rápida reducción del error de entrenamiento según crece el conjunto, mientras que la segunda se basa en el hecho de que, aunque el error de entrenamiento llegue a hacerse cero, el error de test o de generalización puede continuar decreciendo según se añaden más clasificadores al conjunto. Será en estas dos propiedades, analizadas para el caso particular del RA, en las que se centre el contenido de esta sección.

Cuadro 2.1: Pseudocódigo de funcionamiento del algoritmo RA.

1 - Entradas: $\{\mathbf{x}^{(l)}, d^{(l)}\}_{l=1}^{L}$

2 - Inicialización: $D_1(\mathbf{x}^{(l)}) = 1/L \quad \forall l$

3 - Para $t = 1, \ldots, T$

3.1 - Entrenar un clasificador, $o_t(\mathbf{x})$, minimizando la función de coste:

$$C_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2$$

3.2 - Calcular el peso de salida asociado a este clasificador como

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1+\gamma_t}{1-\gamma_t} \right)$$

donde $\gamma_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) o_t(\mathbf{x}^{(l)}) d^{(l)}$

3.3 - Actualizar la función de énfasis para la siguiente iteración:

$$D_{t+1}(\mathbf{x}^{(l)}) = D_t(\mathbf{x}^{(l)}) \exp[-\alpha_t o_t(\mathbf{x}^{(l)}) d^{(l)}]/Z_t$$

siendo $Z_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) \exp[-\alpha_t o_t(\mathbf{x}^{(l)}) d^{(l)}]$

4 - El clasificador final implementa la función

$$f_T(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t o_t(\mathbf{x})$$

siendo, por tanto, $\hat{d}(\mathbf{x}) = \text{sign}\,[f_T(\mathbf{x})]$ la clase estimada para el patrón $\mathbf{x}$.

2.2.1. Convergencia a cero del error de entrenamiento

En 1997 Freund y Schapire (en [Freund y Schapire, 1997]) demostraron que en el caso del “AdaBoost” el error de entrenamiento se reduce a cero a un ritmo exponencial con el número de iteraciones; posteriormente, al proponer la versión del RA, extendieron este resultado a dicho algoritmo [Schapire y Singer, 1999].

La demostración de este resultado para el algoritmo RA se realiza partiendo de la cota sobre E_t^S dada por (2.6) y (2.8) y considerando, además, que los valores de α_t se eligen según las Ecuaciones (2.9)-(2.10); bajo estas premisas, se puede obtener la siguiente cota

del error de entrenamiento en la iteración t-ésima

$$E_t^S = \frac{1}{2L}\sum_{l=1}^{L} \mid \text{sign}\left[f_t(\mathbf{x}^{(l)})\right] - d^{(l)} \mid \leq \prod_{t'=1}^{t}\sqrt{1-\gamma_{t'}^2} \tag{2.11}$$

La demostración de este resultado se proporciona en el Anexo A.2, en el que se analiza un algoritmo más general, que incluye al RA como caso particular.

Teniendo ahora en cuenta que $1 - x \leq \exp(-x)$, para $x > 0$, y definiendo $\gamma^2 = \text{mín}_{t'=1,\ldots t}\{\gamma_{t'}^2\}$, (2.11) puede reescribirse en la forma

$$E_t^S \leq B_t \leq \prod_{t'=1}^{t}\sqrt{1-\gamma_{t'}^2} \leq \exp\left[-\frac{1}{2}\sum_{t'=1}^{t}\gamma_t^2\right] \leq \exp\left[-\frac{t}{2}\gamma^2\right] \tag{2.12}$$

Aunque el valor de γ puede variar en cada iteración, (2.12) sugiere una disminución exponencial del error de entrenamiento según aumenta el número de rondas; de hecho, esta característica ha sido empleada por muchos como argumento para justificar las buenas prestaciones de este algoritmo.

2.2.2. Análisis del error de generalización

Hay que tener en cuenta que el hecho de reducir el error de entrenamiento mediante la incorporación de nuevos clasificadores al conjunto no implica que se esté realizando un diseño sobreajustado, sino todo lo contrario, ya que, simultáneamente a la reducción del error de entrenamiento, se está mejorando el error de generalización, incluso cuando el error de entrenamiento se ha hecho nulo, como se desprende de [Schapire et al., 1998], que justifica su tesis en la combinación de los siguientes resultados:

1. El primer resultado demuestra cómo con un mayor margen de clasificación se mejora el error de generalización del conjunto. Para su explicación, téngase en cuenta que la capacidad de generalización de una red sobre un conjunto de datos generado de forma independiente a partir de una función de densidad de probabilidad $p(x)$ puede formularse en términos del riesgo esperado

$$R[f_T] = E_p\left[\frac{1}{2} \mid \text{sign}\left[f_t(\mathbf{x})\right] - d \mid\right] \tag{2.13}$$

siendo $E_p[\cdot]$ el operador esperanza matemática sobre la función de densidad de probabilidad $p(x)$. Además, tal y como se demuestra en [Schapire y Singer, 1999] para el caso del RA, $R[f_T]$ se mantiene con una probabilidad de al menos $1-\beta$ ($\beta > 0$) por debajo de la cota que se indica:

$$R[f_T] \leq R_T^{\text{margin}}(\theta) + O\left(\frac{1}{\sqrt{L}}\sqrt{\frac{\vartheta \log^2 L/\vartheta}{\theta^2} + \log\frac{1}{\beta}}\right) \tag{2.14}$$

donde $O(\cdot)$ indica que el término al que representa es proporcional a su argumento, L es el tamaño del conjunto de entrenamiento, ϑ es la dimensión VC del espacio de funciones en el que se encuentran los clasificadores [Vapnik y Chervonenkis, 1971], y $R_T^{\text{margin}}(\theta)$ es la fracción de datos de entrenamiento que presentan un margen de clasificación (separación entre la muestra y la frontera de clasificación) menor o igual que $\theta \in (0, 1]$; i.e.,

$$R_T^{\text{margin}}(\theta) = \frac{1}{L}\sum_{l=1}^{L} I\left\{\rho_T(\mathbf{x}^{(l)}) \leq \theta\right\} \tag{2.15}$$

siendo $I\{\cdot\} = 1$ cuando se cumple la condición entre llaves, y 0 en el caso contrario, y donde se ha definido el margen de clasificación de cada muestra , $\rho_T(\mathbf{x}^{(l)})$, como

$$\rho_t(\mathbf{x}^{(l)}) = \frac{f_t(\mathbf{x}^{(l)})d^{(l)}}{\sum_{t'=1}^{t}\alpha_{t'}} \tag{2.16}$$

de modo que $\rho_t(\mathbf{x}^{(l)})$ está comprendido entre -1 y 1, y para las muestras correctamente clasificadas su valor es tanto o más próximo a 1 cuanto mayor es $f_t(\mathbf{x}^{(l)})$ (i.e., conforme mayor es la distancia a la frontera), y de forma similar es próximo a -1 para las muestras mal clasificadas.

La expresión (2.14) permite utilizar el riesgo marginal $R_T^{\text{margin}}(\theta)$, en vez de $R[f_T]$, para analizar la capacidad de generalización del conjunto, ya que indica que para reducir $R[f_T]$ lo importante es conseguir un valor pequeño de $R_T^{\text{margin}}(\theta)$.

2. El segundo resultado indica cómo se comporta el riesgo marginal $R_T^{\text{margin}}(\theta)$, mostrando que el RA no sólo reduce el error de entrenamiento en cada iteración,

sino que además es capaz de reducir $R_T^{\text{margin}}(\theta)$. Este resultado se basa en la existencia de la siguiente cota sobre el riesgo marginal[2]

$$R_T^{\text{margin}}(\theta) \leq \prod_{t=1}^{T}(1+\gamma_t)^{\frac{1+\theta}{2}}(1-\gamma_t)^{\frac{1-\theta}{2}} \tag{2.17}$$

Observando la forma de esta cota se puede afirmar que si los factores que la componen son menores que 1 su valor irá decreciendo según se incorporen clasificadores al conjunto y, consecuentemente, el riesgo marginal, $R_T^{\text{margin}}(\theta)$, tenderá a decrecer.

No obstante, no puede asegurarse que los factores que componen dicha cota van a ser menores que 1, ya que su valor depende tanto de θ como del valor del parámetro de separación de cada clasificador, γ_t; sin embargo, sí que es plausible que se verifique dicha condición cuando se consideran valores pequeños de θ, tal y como puede verse en la Figura 2.2 donde se representa, para distintos valores de θ, el valor del factor t-ésimo en función de γ_t; de hecho, cuando θ tiende a 0 se puede afirmar que los factores son menores que 1 y según se consideran valores más elevados de θ se dificulta que se cumpla tal condición, dependiendo en última instancia del valor del parámetro de separación γ_t que presente el clasificador t-ésimo.

Combinado estos dos resultados, se ve como en cada iteración el RA es capaz de disminuir el riesgo marginal del conjunto y, consecuentemente, reduce el error de generalización; proceso que puede continuar según se añaden clasificadores al conjunto, incluso cuando el error de entrenamiento es nulo.

[2] Véase, para su demostración, [Meir y Rätsch, 2003] o el Anexo A.3 donde partiendo una demostración más general también se obtiene el mismo resultado.

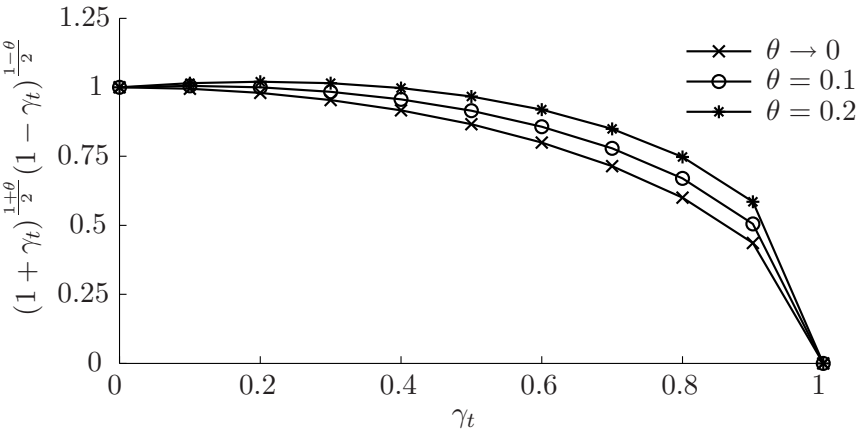

Figura 2.2: Evolución del término t-ésimo de la cota sobre el riesgo marginal en función del parámetro de separación, γ_t, para distintos valores de θ.

2.3. CONCLUSIONES

En este capítulo se ha descrito, en primer lugar, el funcionamiento del algoritmo RA, explicando el procedimiento a seguir para entrenar los clasificadores que lo componen, la actualización de la función de énfasis, $D_t(\cdot)$, que permite que los nuevos clasificadores presten mayor atención a las muestras más erróneas, y la manera de calcular los pesos de la capa de salida $\{\alpha_t\}$ que van asociados a cada clasificador.

A continuación, se han revisado sus propiedades más relevantes, analizando así su comportamiento. Por un lado, se ha mostrado cómo el error de entrenamiento se reduce a un ritmo aproximadamente exponencial; y, por otro lado, se ha analizado la relación existente entre el error de generalización y el margen de clasificación, mostrando cómo el RA es capaz de mejorar la capacidad de generalización del conjunto mediante una reducción del riesgo marginal.

CAPÍTULO 3

"REAL ADABOOST" CON ÉNFASIS MIXTO

Una vez descrito el funcionamiento del algoritmo RA, así como analizadas sus propiedades más relevantes, se comenzará esta sección analizando el mecanismo atencional empleado por el mismo para enfatizar la población, es decir, su función de énfasis. Más concretamente, se demostrará que dicha función realmente se encuentra compuesta de dos términos, uno relacionado con el error cuadrático de las muestras y otro asociado con la distancia de las mismas a la frontera.

Este resultado llevará a proponer un nuevo método de "Boosting". Dicho método modifica la función de énfasis del RA mediante la introducción de un parámetro de mezcla que permite combinar de forma flexible los dos términos de énfasis que el RA combina de manera fija, dando más importancia a uno u otro según convenga en cada problema concreto; además, se verá como, extendiendo el análisis del RA al nuevo algoritmo, se conservan las propiedades que justifican el buen funcionamiento del RA.

3.1. ANÁLISIS DE LA FUNCIÓN DE ÉNFASIS DEL RA

Como ya se ha indicado, el algoritmo RA construye un conjunto de RRNN combinando una serie de clasificadores entrenados con una población de muestras enfatizada de acuerdo con

$$D_{t+1}(\mathbf{x}^{(l)}) = D_t(\mathbf{x}^{(l)}) \exp[-\alpha_t o_t(\mathbf{x}^{(l)}) d^{(l)}]/Z_t \tag{3.1}$$

siendo $D_{t+1}(\mathbf{x}^{(l)})$ la importancia que el clasificador $t+1$-ésimo asigna a la muestra $\mathbf{x}^{(l)}$, $D_1(\mathbf{x}^{(l)}) = 1/L, l = 1, ..., L$, y Z_t una constante de normalización. Mediante esta función de énfasis el algoritmo RA es capaz de dar más importancia a aquellas muestras que han sido clasificadas erróneamente por los clasificadores anteriores, de modo que el nuevo clasificador presta mayor atención a las muestras de más difícil clasificación.

No obstante, el estudio sobre el RA realizado en [Arenas-García et al., 2003] ha puesto de manifiesto que, tras un cierto número de iteraciones, este énfasis acaba centrándose en las muestras más cercanas a la frontera de decisión, lo que sugiere investigar cómo funciona realmente este método de énfasis. Por este motivo, se ha analizado en detalle dicha función de énfasis, concluyendo que, si bien dicha función fue propuesta para poner mayor atención en los patrones más erróneos, en realidad se compone de dos términos: uno que presta atención al error cuadrático de cada dato y otro que se fija en la proximidad del mismo a la frontera.

Para llegar a este resultado, es necesario modificar, en primer lugar, la expresión analítica de la función de énfasis (3.1), reescribiéndola en función de de la salida parcial del conjunto en la ronda t-ésima:

$$\begin{aligned} D_{t+1}(\mathbf{x}^{(l)}) &= \frac{D_t(\mathbf{x}^{(l)}) \exp[-\alpha_t d^{(l)} o_t(\mathbf{x}^{(l)})]}{Z_t} = \frac{\prod_{t'=1}^{t} \exp[-\alpha_{t'} d^{(l)} o_{t'}(\mathbf{x}^{(l)})]}{l \prod_{t'=1}^{t} Z_{t'}} \\ &= \frac{\exp[\sum_{t'=1}^{t} -\alpha_{t'} d^{(l)} o_{t'}(\mathbf{x}^{(l)})]}{Z'_t} \end{aligned} \tag{3.2}$$

donde Z'_t es una nueva constante de normalización cuyo papel es idéntico al de Z_t en (3.1). Recordando ahora que la salida parcial del conjunto para la muestra l-ésima viene dada

por $f_t(\mathbf{x}^{(l)}) = \sum_{t'=1}^{t} \alpha_{t'} o_{t'}(\mathbf{x}^{(l)})$, resulta inmediato escribir

$$D_{t+1}(\mathbf{x}^{(l)}) = \frac{\exp[-f_t(\mathbf{x}^{(l)})d^{(l)}]}{Z'_t} \tag{3.3}$$

Teniendo en cuenta que

$$-2f_t(\mathbf{x}^{(l)})d^{(l)} = [f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2 - f_t^2(\mathbf{x}^{(l)}) - {d^{(l)}}^2$$

es posible reformular (3.3) del siguiente modo:

$$D_{t+1}(\mathbf{x}^{(l)}) = \frac{1}{Z'_t} \exp\left[\frac{[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2}{2} - \frac{f_t^2(\mathbf{x}^{(l)})}{2} - \frac{{d^{(l)}}^2}{2}\right] \tag{3.4}$$

Por último, separando la suma de términos del exponente en producto de exponenciales, y considerando que el último término es constante, ya que ${d^{(l)}}^2 = 1, \forall l$, se obtiene la descomposición deseada para $D_{t+1}(\mathbf{x}^{(l)})$:

$$D_{t+1}(\mathbf{x}^{(l)}) = \frac{1}{\tilde{Z}_t} \exp\left[\frac{[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2}{2}\right] \exp\left[-\frac{f_t^2(\mathbf{x}^{(l)})}{2}\right] \tag{3.5}$$

donde $\tilde{Z}_t$ agrupa todos los términos constantes.

Este resultado evidencia que el énfasis del RA consiste en la combinación fija de dos factores, representando cada uno de ellos un tipo diferente de énfasis:

1. **Énfasis sobre el error cuadrático**: el primer factor, correspondiente al término

   $$\exp\left[\frac{[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2}{2}\right]$$

   aumenta conforme crece el error cuadrático que presenta cada muestra.

2. **Énfasis sobre la frontera de clasificación**: el segundo factor, dado por

   $$\exp\left[-\frac{f_t^2(\mathbf{x}^{(l)})}{2}\right]$$

   es mayor para las muestras cercanas a la frontera de clasificación, que de ahora en adelante serán también llamadas "críticas". Nótese que para la iteración $t+1$ se está midiendo la cercanía del dato l-ésimo a la frontera de clasificación mediante el valor de $f_t^2(\mathbf{x}^{(l)})$, por lo que se está considerando que en cada iteración la frontera de clasificación es aquella dada por el conjunto en la ronda anterior, es decir, la solución de $f_t(\mathbf{x}) = 0$.

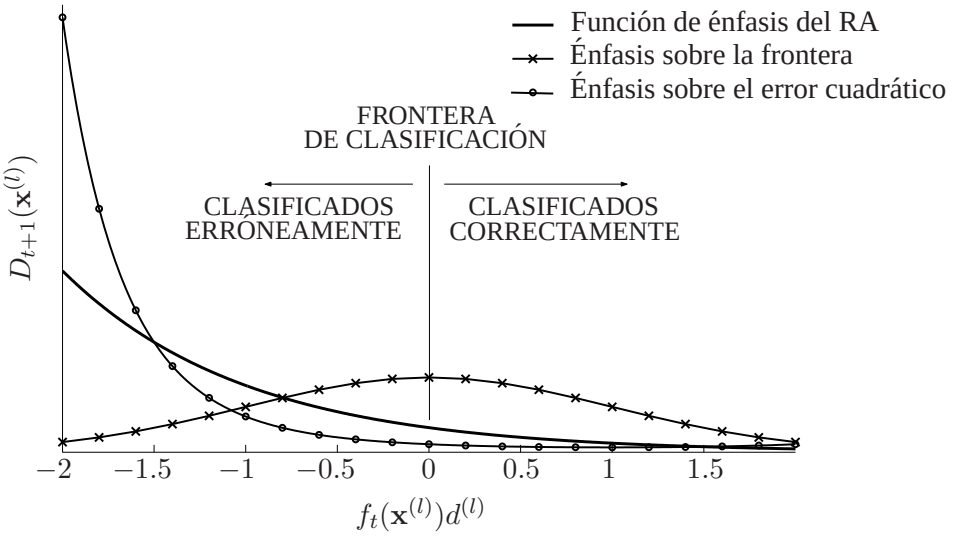

Figura 3.1: Influencia de los distintos tipos de énfasis sobre cada tipo de datos.

En la Figura 3.1 se indica la importancia que la función de énfasis del RA asigna a los distintos tipos de muestras, así como la atención que les presta cada tipo de énfasis (énfasis sobre el error cuadrático y énfasis sobre la frontera). Nótese que, a modo ilustrativo, en la figura se ha limitado el rango de valores de $f_t(\mathbf{x}^{(l)})d^{(l)}$ al intervalo $[-2, 2]$, pero realmente el producto $f_t(\mathbf{x}^{(l)})d^{(l)}$ puede tomar cualquier valor en $\Re$.

3.2. FUNCIÓN DE ÉNFASIS MIXTO

La descomposición de la función de énfasis presentada en (3.5) muestra que el RA combina dos tipos de énfasis de forma fija, ponderando por igual la importancia de cada uno de estos términos; esto lleva a preguntarse si sería preferible disponer de un ajuste flexible que permita dar más importancia a un término u otro. De hecho, estudios anteriores [Rätsch et al., 2001] han demostrado que el énfasis del RA suele proporcionar soluciones sobreajustadas cuando entre las muestras de entrenamiento hay presentes “outliers”, lo que implica que, en ocasiones, la manera de enfatizar la población del RA presente serios inconvenientes.

Por este motivo, en esta Tesis Doctoral se propone una nueva función de énfasis consistente en una combinación flexible de los dos términos de énfasis, sobre el error cuadrático y sobre proximidad a la frontera. Para ello, se incorpora un grado de libertad adicional a la combinación fija empleada por el RA en (3.5), de modo que

$$\boxed{D_{\lambda,t+1}(\mathbf{x}^{(l)}) = \frac{1}{Z_{\lambda,t}} \exp\left\{\lambda[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1-\lambda) f_t^2(\mathbf{x}^{(l)})\right\}} \tag{3.6}$$

donde λ ($0 \leq \lambda \leq 1$) es un parámetro de mezcla empleado para ponderar la atención que el conjunto presta a las muestras "críticas" y a las que presentan un elevado error cuadrático, mientras que $Z_{\lambda,t}$ es una constante de normalización que asegura que $\sum_{l=1}^{L} D_{\lambda,t+1}(\mathbf{x}^{(l)}) = 1$, y se calcula en cada iteración como

$$Z_{\lambda,t} = \sum_{l=1}^{L} \exp\left\{\lambda[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1-\lambda) f_t^2(\mathbf{x}^{(l)})\right\} \tag{3.7}$$

De este modo, la nueva función de énfasis $D_{\lambda,t+1}$ permite controlar la atención concedida a los distintos tipos de datos según el valor del parámetro de mezcla seleccionado, destacando tres casos particulares asociados a tres valores de λ:

- $\lambda = 0$: la función de énfasis $D_{\lambda,t+1}$ se reduce al término de énfasis en la frontera, de manera que los sucesivos clasificadores centran su atención únicamente en las muestras "críticas".

- $\lambda = 0.5$: en este caso se concede igual importancia a ambos términos de énfasis, obteniéndose la función de énfasis del RA (compárese (3.6) con (3.5) para este valor de λ).

- $\lambda = 1$: en este caso, la función de énfasis se fija únicamente en el error cuadrático de los datos, dando mayor importancia a aquéllos que presentan un mayor error cuadrático.

Aunque el empleo de criterios para asignar más o menos importancia a las muestras durante el entrenamiento de una red neuronal no es nuevo, ya que es posible encontrar otros

trabajos previos sobre métodos que prestan atención al error de cada muestra, como es el caso de [Cachin, 1994, Kung y Taur, 1995, Munro, 1992], así como métodos que se centran en la proximidad de las muestras a la frontera de clasificación (véanse, por ejemplo, los métodos propuestos en [Choi y Rockett, 2002, Hart, 1968, Wann et al., 1990]), cabe destacar que la función de énfasis propuesta posibilita una combinación flexible de ambos criterios, permitiendo así emplear criterios atencionales más ajustados a las peculiaridades de cada problema de clasificación concreto.

3.3. CONSTRUCCIÓN DE CONJUNTOS DE RRNN CON ÉNFASIS MIXTO

A la hora de diseñar un conjunto de RRNN empleando esta nueva función de énfasis, se puede elegir un valor del parámetro de mezcla y seguir el mismo procedimiento empleado por el RA, es decir, entrenar iterativamente un conjunto de clasificadores con una población de datos remuestreada, aunque ahora empleando la función (3.6); para ello, en cada iteración se entrenará un clasificador base de modo que minimice el error cuadrático medio de los datos ponderado por el término de énfasis dependiente del valor de λ, es decir, se empleará la función de coste:

$$C_{\lambda,t} = \sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2 \tag{3.8}$$

donde $D_{\lambda,t}(\cdot)$ deberá actualizarse en cada iteración según (3.6).

Una vez entrenado cada clasificador base, se le asignará un peso de salida de modo que se minimice, al igual que en el RA, la cota del error de entrenamiento presentada en (2.8). Tal y como se demuestra en el Apéndice A.1, esto se consigue seleccionando el correspondiente peso de salida, α_t, de acuerdo con

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 + \delta_t}{1 - \delta_t} \right) \tag{3.9}$$

donde δ_t viene dado por la expresión

$$\delta_t = \frac{1}{LB_{t-1}} \sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]o_t(\mathbf{x}^{(l)})d^{(l)} \tag{3.10}$$

Esta expresión es equivalente a la de γ_t, ya que, para el algoritmo RA ($\lambda = 0.5$), $D_t(\mathbf{x}^{(l)}) \propto \exp[-f_t(\mathbf{x}^{(l)})d^{(l)}]$; sin embargo, hay una clara diferencia entre sus interpretaciones:

- La expresión de δ_t mide la correlación entre las salidas del clasificador base para cada dato y las correspondientes etiquetas, pero ponderada por la contribución de cada dato a B_{t-1} (definido en (2.6)).
- La expresión para γ_t mide la misma correlación, pero en este caso se considera que cada término de la correlación está ponderando por la importancia (énfasis) del patrón asociado.

Evidentemente, ambos puntos de vista son similares cuando $D_{\lambda,t}(\mathbf{x}^{(l)}) = D_t(\mathbf{x}^{(l)})$, i.e., cuando $\lambda = 0.5$. Por lo tanto, cuando se emplea la función de énfasis mixto se prefiere emplear la expresión (3.10), ya que tiene una interpretación clara para cualquier valor de λ; por este motivo, se denotará δ_t como parámetro generalizado de separación del clasificador ("generalized edge").

La aplicación de la función de énfasis (3.6) para un valor concreto de λ, junto con la función de coste (3.8) para entrenar los clasificadores y la selección de los valores α_t mediante (3.9)-(3.10), da origen a un nuevo algoritmo para la construcción de conjuntos mediante "Boosting", al que se ha llamado RA con énfasis ponderados o mixtos y que, de ahora en adelante, se va a denominar RA-we ("RA with weighted emphasis"). El pseudocódigo que describe el funcionamiento de este algoritmo se ofrece en el Cuadro 3.1. Nótese que no existe diferencia alguna entre el RA y el RA-we durante sus fases de clasificación.

Cuadro 3.1: Pseudocódigo de funcionamiento del algoritmo RA-we.

1 - Entradas: $\{\mathbf{x}^{(l)}, d^{(l)}\}_{l=1}^{L}$, λ

2 - Inicializa: $D_{\lambda,1}(\mathbf{x}^{(l)}) = 1/L, \forall l$ y $f_0(\mathbf{x}^{(l)}) = 0, \forall l$

3 - Para $t = 1, \ldots, T$

3.1 - Entrenar un clasificador, $o_t(\mathbf{x})$, que minimice la función de coste:

$$C_{\lambda,t} = \sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2$$

3.2 - Calcular el parámetro generalizado de separación del clasificador como

$$\delta_t = \frac{1}{LB_{t-1}} \sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]o_t(\mathbf{x}^{(l)})d^{(l)}$$

donde $B_{t-1} = \frac{1}{L} \sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]$

3.3 - Calcular el peso de salida asociado a este clasificador como

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1+\delta_t}{1-\delta_t}\right)$$

3.4 - Actualizar la función de énfasis para la siguiente iteración:

$$D_{\lambda,t+1}(\mathbf{x}^{(l)}) = \frac{1}{Z_{\lambda,t}} \exp\left\{\lambda[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1-\lambda)f_t^2(\mathbf{x}^{(l)})\right\}$$

siendo $Z_{\lambda,t} = \sum_{l=1}^{L} \exp\left\{\lambda[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1-\lambda)f_t^2(\mathbf{x}^{(l)})\right\}$

4 - El clasificador final implementa la función

$$f_T(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t o_t(\mathbf{x})$$

siendo, por tanto, $\hat{d}(\mathbf{x}) = \text{sign}\,[f_T(\mathbf{x})]$ la clase estimada para el patrón $\mathbf{x}$.

3.4. ANÁLISIS DE LAS PROPIEDADES DEL RA-WE

En el Capítulo 2 se consideraron dos propiedades importantes del algoritmo RA; la primera de ellas muestra que el error de entrenamiento decrece a un ritmo exponencial según aumenta el número de clasificadores que forman el conjunto, mientras que la segunda explica la mejora de la capacidad de generalización del conjunto con el número de iteraciones, incluso cuando el error de entrenamiento se ha hecho nulo. Los dos siguientes apartados presentan una extensión de dichos resultados al RA-we, evidenciando que el nuevo algoritmo también disfruta ambas propiedades.

3.4.1. Convergencia del error de entrenamiento

Tal y como se presentó en el Apartado 2.2.1, Schapire y Singer [Schapire y Singer, 1999] establecieron una cota para el error del entrenamiento del RA que permite demostrar que dicho error decrece a un ritmo aproximadamente exponencial con el número de iteraciones. Una cota similar puede establecerse para el algoritmo RA-we, tal y como se demuestra en el Anexo A.2,

$$E_t^S = \frac{1}{2L}\sum_{l=1}^{L} \mid \text{sign}\left[f_t(\mathbf{x}^{(l)})\right] - d^{(l)} \mid \leq B_t \leq \prod_{t'=1}^{t}\sqrt{1-\delta_{t'}^2} \leq \exp\left[-\frac{\delta^2}{2}t\right] \qquad (3.11)$$

donde se introduce δ^2 como $\delta^2 = \text{mín}_{t'=1,\ldots t}\{\delta_{t'}^2\}$. Y, del mismo modo que se hacía para el RA, se puede afirmar que el error de entrenamiento decrece de manera aproximadamente exponencial con el número de rondas.

3.4.2. Análisis del error de generalización

En el Apartado 2.2.2 se justificaba por qué el error de test del RA puede continuar decreciendo según se van añadiendo más clasificadores al conjunto, incluso cuando el error de entrenamiento se ha hecho nulo, presentando para ello dos resultados intermedios. A continuación, se extienden ambos resultados para el algoritmo RA-we:

1. El primero de estos resultados se basa en la cota (2.14) riesgo esperado del clasificador, que se repite aquí por conveniencia,

$$R[f_T] \leq R_T^{\text{margin}}(\theta) + O\left(\frac{1}{\sqrt{L}}\sqrt{\frac{\vartheta \log^2 L/\vartheta}{\theta^2} + \log\frac{1}{\beta}}\right) \tag{3.12}$$

   la cual permite utilizar el riesgo marginal $R_T^{\text{margin}}(\theta) = \sum_{l=1}^{L} I\left\{\rho_T(\mathbf{x}^{(l)}) \leq \theta\right\}$, en vez de $R[f_T]$, para analizar la capacidad de generalización del conjunto.

   La validez de (3.12) radica en el hecho de que la salida global del conjunto se obtiene mediante la combinación lineal de un conjunto de clasificadores, $\{o_t\}_{t=1}^{T}$ ponderados por un conjunto de pesos no negativos, $\{\alpha_t\}_{t=1}^{T}$, (véase [Schapire y Singer, 1999] para un análisis más detallado de esta cota); dado que en el algoritmo propuesto dichas condiciones siguen verificándose, la expresión (3.12) también es válida para el caso del algoritmo RA-we.

2. El segundo de los resultados permitía afirmar que el RA no sólo reduce el error de entrenamiento en cada iteración, sino que además reduce el riesgo marginal. Para extender este resultado al RA-we, en el Anexo A.3 se muestra la existencia de una cota equivalente cuando el parámetro de separación empleado es δ_t en vez de γ_t, permitiendo así establecer la siguiente cota sobre $R_T^{\text{margin}}(\theta)$

$$R_T^{\text{margin}}(\theta) \leq \prod_{t=1}^{T}(1+\delta_t)^{\frac{1+\theta}{2}}(1-\delta_t)^{\frac{1-\theta}{2}} \tag{3.13}$$

   que permite afirmar que, cuando los factores que la componen son menores que 1, el $R_T^{\text{margin}}(\theta)$ tiende a decrecer con el número de rondas (especialmente para valores pequeños de θ).

La combinación de ambos resultados permite justificar que el RA-we puede reducir el riesgo marginal del conjunto, permitiendo así mejorar su capacidad de generalización, incluso cuando el error de entrenamiento se haya hecho nulo.

Nótese, además, que si $\delta_t < \gamma_t \ \forall t$ la cota obtenida por el RA-we será más favorable que la presentada por el RA. Si bien no se puede afirmar que éste sea el caso, la posibilidad

de explorar el valor de λ sugiere que es posible que valores $\lambda \neq 0.5$ ofrezcan, al menos en algunos casos, cotas más ajustadas.

3.5. INFLUENCIA DEL ÉNFASIS MIXTO EN EL ENTRENAMIENTO DE LOS CLASIFICADORES BASE

A la hora de construir un conjunto tipo RA, el diseño y/o entrenamiento de los clasificadores base juega un papel muy importante, por lo que dicho diseño y/o entrenamiento debe realizarse de modo que aporte las mejores prestaciones desde un punto de vista global. Por este motivo, han surgido trabajos [Rätsch et al., 2002] que analizan este aspecto, en los que se propone entrenar los clasificadores base para maximizar su parámetro de separación (2.10), y así minimizar directamente B_t, que es lo que realmente importa.

En esta sección se va a analizar la función de coste empleada en los algoritmos RA y RA-we, y se verá cómo en ambos casos existe una conexión con el parámetro de separación del clasificador (γ_t o δ_t, según el caso), mostrando, además, cómo el hecho de emplear la función de énfasis mixto (3.6) permite enfatizar la influencia que las distintos tipos de muestras tienen sobre dicho parámetro.

Considerando en primer lugar la función de coste empleada por el RA para el entrenamiento de los clasificadores base

$$C_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2 \tag{3.14}$$

se puede demostrar que la minimización de C_t es equivalente a maximizar una versión regularizada del parámetro de separación. Esta equivalencia se comprueba al desarrollar

el término cuadrático de la Ecuación (3.14):

$$\begin{aligned} C_t &= \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) \left[{d^{(l)}}^2 + o_t^2(\mathbf{x}^{(l)}) - 2o_t(\mathbf{x}^{(l)})d_l \right] \\ &= \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) {d^{(l)}}^2 + \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) o_t^2(\mathbf{x}^{(l)}) - 2\sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) o_t(\mathbf{x}^{(l)}) d^{(l)} \end{aligned} \quad (3.15)$$

Teniendo en cuenta que el primer término es constante ($\sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) {d^{(l)}}^2 = 1$) y, por lo tanto, irrelevante para el proceso de minimización, se obtiene la siguiente función de coste equivalente

$$\begin{aligned} C_t &\stackrel{c}{=} \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) o_t^2(\mathbf{x}^{(l)}) - 2\sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) o_t(\mathbf{x}^{(l)}) d^{(l)} \\ &= \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)}) o_t^2(\mathbf{x}^{(l)}) - 2\gamma_t \end{aligned} \quad (3.16)$$

donde la notación $a \stackrel{c}{=} b$ se emplea para indicar que a y b sólo difieren en una constante.

La minimización de C_t es, por tanto, equivalente a la maximización del parámetro de separación del clasificador, salvo por la aparición de un término de regularización que penaliza los valores elevados de las salidas. Dicha componente de penalización favorece el entrenamiento de los clasificadores base posibilitando el empleo de técnicas de descenso por gradiente.

En segundo lugar, considérese la función de coste empleada por RA-we,

$$C_{\lambda,t} = \sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2$$

y descompóngase de forma equivalente a como se hizo para C_t:

$$C_{\lambda,t} \stackrel{c}{=} \sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}^{(l)}) o_t^2(\mathbf{x}^{(l)}) - 2\sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}_l) o_t(\mathbf{x}_l) d^{(l)} \quad (3.17)$$

que puede escribirse de forma más conveniente como

$$C_{\lambda,t} \stackrel{c}{=} \sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}^{(l)}) o_t^2(\mathbf{x}^{(l)}) - 2C_{\lambda,t}^{\delta} \quad (3.18)$$

de manera que la minimización de $C_{\lambda,t}$ puede interpretarse como la maximización de $C^{\delta}_{\lambda,t}$ más un término de regularización que penaliza las salidas elevadas. Al contrario que el caso del RA, $C^{\delta}_{\lambda,t}$ no corresponde al parámetro de separación del RA-we, sino que es una versión enfatizada del mismo. Efectivamente, en el Anexo A.4 se demuestra que $C^{\delta}_{\lambda,t}$ puede calcularse como

$$C^{\delta}_{\lambda,t} = \frac{\tilde{Z}}{LB_{t-1}} \sum_{l=1}^{L} G_{\lambda,t}(\mathbf{x}^{(l)}) \exp\left[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}\right] o_t(\mathbf{x}^{(l)})d^{(l)} \tag{3.19}$$

donde $\tilde{Z}$ es una constante, irrelevante en el proceso de minimización, y $G_{\lambda,t}(\mathbf{x}^{(l)})$ es una nueva función de énfasis definida como

$$G_{\lambda,t}(\mathbf{x}^{(l)}) = \frac{1}{Z_{G,t}} \exp\left\{2(\lambda - 0.5)\left[f_{t-1}(\mathbf{x}^{(l)}) - \frac{d^{(l)}}{2}\right]^2\right\} \tag{3.20}$$

siendo $Z_{G,t}$ un factor de normalización que asegura que $\sum_{l=1}^{L} G_{\lambda,t}(\mathbf{x}^{(l)}) = 1$.

Comparando (3.19) y (3.10), es posible concluir que en el RA-we los clasificadores base se entrenan para maximizar una versión enfatizada (y regularizada) del parámetro de separación asociado a cada clasificador, donde diferentes valores de λ dan lugar a distintos tipos de énfasis; por ejemplo:

- Si $\lambda = 0$ (recuérdese que en este caso la función de énfasis mixto, $D_{\lambda,t}$, sólo enfatiza las muestras críticas), se obtiene que

$$G_{\lambda,t}(\mathbf{x}^{(l)})|_{\lambda=0} = \frac{1}{Z_{G,t}} \exp\left\{-\left[f_{t-1}(\mathbf{x}^{(l)}) - \frac{d^{(l)}}{2}\right]^2\right\} \tag{3.21}$$

  lo que indica, desde el punto de vista de maximización de δ_t, que las muestras correctamente clasificadas con valores de salida cercanos a 0.5 ó -0.5 se consideran más relevantes durante el entrenamiento del clasificador para la maximización de δ_t.

- Cuando se considera el caso particular del RA, $\lambda = 0.5$, resulta que $G_{\lambda,t}(\mathbf{x}^{(l)})$ es una distribución uniforme, y por lo tanto se entrena el clasificador considerando que todos los datos deben influir por igual en la maximización de δ_t.

- Por último, si $\lambda = 1$ ($D_{\lambda,t}$ se centra en los datos con un elevado error cuadrático), resulta que

$$G_{\lambda,t}(\mathbf{x}^{(l)})|_{\lambda=1} = \frac{1}{Z_{G,t}} \exp\left\{\left[f_{t-1}(\mathbf{x}^{(l)}) - \frac{d^{(l)}}{2}\right]^2\right\} \tag{3.22}$$

  y son las muestras que presentan un mayor error cuadrático las que se consideran más importantes durante el entrenamiento del clasificador para maximizar δ_t.

Por lo tanto, si se considera el entrenamiento de los clasificadores base desde el punto de vista de maximización del parámetro de separación, se puede afirmar que el RA concede a todos los datos la misma influencia, mientras que el RA-we, por el hecho de emplear un tipo de énfasis ajustable, puede considerar cierto tipo de muestras más relevantes para la maximización de dicho parámetro, adaptando así la construcción del conjunto a las características del problema de clasificación a resolver.

3.6. SELECCIÓN DE λ POR VALIDACIÓN CRUZADA

Como es fácilmente deducible, y tal y como se analiza en [Gómez-Verdejo et al., 2005] y se mostrará en el Capítulo 6 de esta Tesis Doctoral, las prestaciones del algoritmo RA-we están completamente ligadas al valor del parámetro de mezcla, pudiéndose obtener mejoras considerables frente al RA para ciertos valores de λ; sin embargo, el valor óptimo del parámetro de mezcla está ligado a las características de cada problema de clasificación, por lo que su selección determinará en última instancia las prestaciones del conjunto en un problema concreto.

Una posibilidad para seleccionar un parámetro a fijar es el empleo de un proceso de validación cruzada o "Cross Validation" (CV) [Bishop, 1995]; dicho método consiste en dividir el conjunto de datos de entrenamiento en K particiones y emplear de manera secuencial $K - N$ particiones para el entrenamiento, utilizando distintos valores del parámetro a determinar, y las N restantes para validar el resultado, fijando el parámetro libre a aquel

valor que resulta en un error de validación menor.

La aplicación de este procedimiento para la selección de λ ha dado lugar a una versión particular del algoritmo RA-we, a la que se ha denominado CV RA-we, y que, tal y como se muestra en [Gómez-Verdejo et al., 2006] y se verá en detalle en el Capítulo 6, proporciona ventajas frente al RA en términos de reducción de la tasa de error, aceleración de la velocidad de convergencia y robustez frente al sobreajuste. No obstante lo anterior, se ha comprobado que en la mayoría de los casos el uso de validación cruzada no determina el valor óptimo de λ, por lo que el CV RA-we no aprovecha al máximo las ventajas del énfasis mixto. Por este motivo, se han propuesto dos esquemas alternativos para la selección de λ que no precisan del proceso de CV. Dichos esquemas serán objeto de estudio en los dos capítulos siguientes de esta Tesis.

3.7. CONCLUSIONES

En este capítulo se ha analizado la función de énfasis empleada por el RA, mostrando que dicha función está compuesta de dos términos: el primero de ellos enfatiza los datos que presentan un mayor error cuadrático, mientras que el segundo enfatiza los datos críticos (aquellos más próximos a la frontera de clasificación).

Este resultado ha llevado a proponer una función de énfasis mixto ajustable, consistente en la combinación flexible de los dos términos de énfasis mediante un parámetro de mezcla que permite asignar más o menos importancia a las muestras que presentan un mayor error cuadrático o una menor distancia a la frontera de decisión. Junto con esta nueva función de énfasis se ha propuesto una nueva función de coste para el entrenamiento de los clasificadores base, así como un nuevo criterio para seleccionar los pesos de salida asociados a cada clasificador, surgiendo, de este modo, un nuevo algoritmo de "Boosting" al que se ha denominado RA-we ("RA with weighted emphasis").

Teóricamente, el algoritmo RA-we sigue disfrutando de las propiedades más atractivas del RA, pero a diferencia del RA y debido a la introducción de un tipo de énfasis ajustable,

el RA-we puede adaptar la construcción del conjunto a las características de cada problema de clasificación.

Por último, se ha presentado un primer método, al que se ha denominado CV RA-we, que utiliza CV para la selección del parámetro de mezcla. A pesar de sus ventajas frente al RA básico, el uso de CV no permite al RA-we explotar al máximo las posibilidades de la nueva función de énfasis mixto, ya que sólo se utiliza una parte de la información contenida en el conjunto de entrenamiento para ajustar el parámetro de mezcla.

En los dos próximos capítulos de esta Tesis Doctoral se profundizará en la propuesta de métodos para la selección automática de λ. Estos métodos permiten emplear todos los patrones de entrenamiento disponibles para el ajuste de λ, lo que resultará en algoritmos de construcción de conjuntos más potentes.

CAPÍTULO 4

COMITÉS DE CONJUNTOS RA-WE

Como se ha explicado, el empleo de un énfasis mixto en la construcción de conjuntos puede proporcionar considerables ventajas frente al RA básico, pero, para ello, es necesario emplear un método que permita una adecuada selección del parámetro de mezcla, λ. Además, también se ha discutido que el empleo para tal fin de un proceso de validación cruzada (CV) no parece ser la mejor opción, ya que no es capaz de aprovechar al máximo las ventajas que este énfasis mixto es capaz de aportar. Por este motivo, en este capítulo y en el siguiente se van a presentar esquemas alternativos que permiten una mejor explotación de estas posibilidades y que no precisan del proceso de CV.

La primera de estas alternativas no opta por buscar el mejor valor del parámetro de mezcla, sino que aprovecha la diversidad existente entre conjuntos construidos con distintos valores del parámetro de mezcla para construir un comité de conjuntos RA-we; así el comité resultante puede conseguir mejores resultados que los presentados individualmente por cualquiera de los conjuntos RA-we. Será esta primera alternativa en la que se centrará el contenido de este capítulo, proponiendo una serie de métodos que permiten combinar conjuntos RA-we y formar así comités de RA-we; además, junto con la propuesta de estos métodos, se presentará una técnica para seleccionar entre todos los conjuntos

RA-we disponibles aquéllos que resulten más adecuados para constituir el comité.

Para finalizar el capítulo, se detallará un nuevo método, sencillo y eficiente, que permite reducir el tiempo de cómputo que requiere la fase de clasificación de estos comités.

4.1. CONSTRUCCIÓN DE COMITÉS DE RA-WE

Como se sabe, un comité de RRNN consiste en una combinación de una serie de redes de tal manera que (idealmente) la red global supere en prestaciones a cualquiera de las redes aisladas; para ello es necesario, por un lado, diseñar estas redes de modo que generalicen de forma diferente [Krogh y Vedelsby, 1995] y, por otro lado, seleccionar un criterio adecuado para combinar sus salidas.

Siguiendo esta idea, se va a aprovechar la diversidad existente entre conjuntos RA-we entrenados con distintos tipos de énfasis (distintos valores de λ) para construir comités de conjuntos RA-we. Dado que el procedimiento para entrenar las redes RA-we ya está fijado, el objetivo será encontrar el método más adecuado para combinar sus salidas.

A lo largo de este capítulo se va a trabajar con una combinación de "segundo nivel". Por ello, de cara a evitar confusiones, se será más sistemático con la nomenclatura, reservando el término *conjunto* para las redes RA-we (primer nivel) y el término *comité* para la red resultante de la agrupación de conjuntos RA-we (segundo nivel jerárquico).

4.1.1. Combinación de las salidas de conjuntos RA-we

Entre los diferentes métodos que se pueden encontrar en la literatura para la construcción de comités de RRNN (véase [Kuncheva, 2004], donde se recogen una amplia gama de opciones), destacan dos tipos de esquemas: las combinaciones lineales y los esquemas de voto.

Entre las combinaciones lineales se encuentra, como esquema más directo, la ponderación equitativa de las salidas de todas las redes [Breiman, 1996], lo que equivale a

calcular la salida global como el valor promedio; aunque éste es el criterio más sencillo, presenta el inconveniente de dar la misma importancia a todos las redes, sin tener en cuenta que algunas pueden presentar mejores prestaciones que otras. Por este motivo, suelen preferirse técnicas de optimización basadas en la minimización del error cuadrático medio ("Mean Square Error", MSE) de la salida global, que permiten ajustar el peso o importancia otorgada a cada red teniendo en cuenta sus prestaciones individuales [Hashem, 1995, Meir, 1995]. Junto con estos métodos se pueden encontrar otro tipo de técnicas que también permiten ajustar el peso de cada red, pero que emplean criterios diferentes a la minimización del MSE; entre ellos destacan los métodos Bayesianos, el método de Dempster-Shafer y los métodos de regresión logística (véase [Bahler y Navarro, 2000], donde se describen estas técnicas).

Por otro lado, las técnicas de voto [Auda et al., 1995, Battiti y Colla, 1995] deciden si un dato pertenece a una clase según la proporción de redes individuales que están de acuerdo en asignar tal clase; para el caso de clasificación binaria suele emplearse la votación por mayoría, en la que la salida global consiste en asignar la clase que la mayoría de los clasificadores dan por correcta. Aunque este método es sencillo, al igual que la selección del valor promedio de las salidas tiene el problema de que todos los clasificadores influyen igualmente en la salida final, y no se considera la posibilidad que algunos puedan presentar mejores prestaciones que otros.

De entre estos criterios, en este capítulo se consideran comités de redes RA-we que utilizan combinaciones lineales basadas en la minimización del MSE, así como esquemas de votación generalizada. Además, se considera una alternativa adicional basada en el criterio seguido por el RA-we para la selección de los pesos de salida. Aunque esta última alternativa es típica de esquemas de consorcio, y normalmente no se emplea para la construcción de comités, es plausible que aporte ventajas frente a los métodos anteriores dadas las características de las redes que se están considerando.

Antes de pasar a describir los diversos esquemas mencionados, se va a introducir la nomenclatura que se utilizará para todos ellos. Considérese que se tienen J conjuntos RA-we, cada uno de ellos entrenado con un valor diferente del parámetro de mezcla de entre el

conjunto de valores $\{\lambda^{(1)}, \ldots, \lambda^{(J)}\}$; por lo que se va a denotar a cada una de sus salidas como $f^{(j)}(\mathbf{x})$, $j = 1, \ldots, J$, siendo la función j-ésima la correspondiente al parámetro de mezcla $\lambda^{(j)}$. El objetivo será construir un comité combinando estas salidas mediante un conjunto de pesos $\{w_1, \ldots, w_J\}$, que se ajustarán empleando el conjunto de datos de entrenamiento

$$S = \{(\mathbf{x}^{(1)}, d^{(1)}), \ldots, (\mathbf{x}^{(L)}, d^{(L)})\}, \qquad d^{(l)} \in \{-1, +1\}, \quad l = 1, \ldots, L$$

A continuación se enumeran los distintos métodos de construcción de comités considerados, explicando cómo se lleva a cabo la selección de los pesos $\{w_1, \ldots, w_J\}$ para cada uno de ellos.

1. **Comité basado en combinación lineal de mínimo MSE**

   En este primer método, la salida final del comité, $F_{\text{lin}}(\mathbf{x})$, viene dada por

   $$F_{\text{lin}}(\mathbf{x}) = \sum_{j=1}^{J} w_j f^{(j)}(\mathbf{x}) \tag{4.1}$$

   y el conjunto de pesos $\{w_1, \ldots, w_J\}$ se ajusta de modo que se minimice el MSE sobre el conjunto de datos de entrenamiento S, es decir,

   $$\text{MSE} = \frac{1}{L} \sum_{l=1}^{L} \left[F_{\text{lin}}(\mathbf{x}^{(l)}) - d^{(l)}\right]^2 \tag{4.2}$$

   Si se emplea notación vectorial, y se denota al conjunto de pesos como el vector $\mathbf{w} = [w_1, \ldots, w_J]^T$, se puede obtener directamente el valor de $\mathbf{w}$ que minimiza (4.2) mediante

   $$\mathbf{w} = \mathbf{F}^{\dagger}\mathbf{d} \tag{4.3}$$

   donde $\mathbf{d}$ es el vector formado por las etiquetas de todos los datos de entrenamiento, $\mathbf{d} = [d^{(1)}, \ldots, d^{(L)}]^T$, y $\mathbf{F}$ es la matriz formada por las salidas de las J redes para cada uno de los datos del conjunto de entrenamiento

   $$\mathbf{F} = \begin{bmatrix} f^{(1)}(\mathbf{x}^{(1)}) & \ldots & f^{(J)}(\mathbf{x}^{(1)}) \\ \vdots & \ddots & \vdots \\ f^{(1)}(\mathbf{x}^{(L)}) & \ldots & f^{(J)}(\mathbf{x}^{(L)}) \end{bmatrix} \tag{4.4}$$

Finalmente, en (4.3) se ha utilizado $\mathbf{F}^{\dagger}$ para denotar la inversa generalizada de Moore-Penrose de la matriz $\mathbf{F}$, que se obtiene según: $\mathbf{F}^{\dagger} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T$ (supuesto que $\mathbf{F}^T\mathbf{F}$ es una matriz invertible, lo que es habitual; en caso contrario, se puede recurrir a su regularización o a técnicas de descomposición en valores singulares).

2. **Comité basado en combinación lineal con función de activación tipo tangente hiperbólica**

   Para limitar el rango de la salida del comité al intervalo $[-1, 1]$, se puede incorporar a la salida del mismo una función de activación de tipo tangente hiperbólica, de modo que la salida final del comité venga dada por

$$F_{\text{th}}(\mathbf{x}) = \tanh\left(\sum_{j=1}^{J} w_j f^{(j)}(\mathbf{x})\right) \tag{4.5}$$

   Aunque en este caso también se opta por seleccionar el conjunto de pesos $\{w_1, \ldots, w_J\}$ que minimiza el MSE, no existe una solución cerrada para el cálculo de $\{w_1, \ldots, w_J\}$, y es necesario recurrir a algoritmos de búsqueda.

   Entre los múltiples algoritmos de búsqueda que existen [Bishop, 1995, Haykin, 1999], se ha optado por emplear el bien conocido algoritmo de descenso por gradiente estocástico: tras inicializar aleatoriamente el conjunto de pesos $\{w_1, \ldots, w_J\}$, se actualizan iterativamente sus valores en dirección contraria a la estimación instantánea del gradiente de la función de error, de modo que el vector de pesos se acerque progresivamente al mínimo de la función de coste. Cuando se minimiza el MSE, y dado que la salida del sistema viene dada por (4.5), la regla de actualización de cada peso es:

$$w_j^{k+1} = w_j^k - \eta\left[1 - F_{\text{th}}^2(\mathbf{x}^{(l)})\right]\left[d^{(l)} - F_{\text{th}}(\mathbf{x}^{(l)})\right] f^{(j)}(\mathbf{x}^{(l)}), \quad j = 1, \ldots, J \tag{4.6}$$

   donde k es la iteración y η la constante de aprendizaje del algoritmo, que, como se sabe, marca un compromiso entre la velocidad de convergencia y el error residual de la solución. Tras un número suficiente de iteraciones, se llegará a un mínimo del MSE, o, al menos, se estará lo suficientemente cerca de él y la regla de aprendizaje

se detendrá, indicando el conjunto de valores $\{w_1, \ldots, w_J\}$ que proporcionan un mínimo local de la función de coste o, en el mejor de los casos, su mínimo global.

3. **Comité basado en esquema de voto generalizado**

   Las técnicas de voto por mayoría aplicadas a la resolución de problemas binarios deciden la clase a la que pertenece un dato como aquélla en la que la mayoría de redes están de acuerdo. Aunque este método es de aplicación inmediata y no requiere ningún tipo de información sobre las prestaciones individuales de los clasificadores, tiene el problema de considerar igualmente válido el criterio de todas las redes. Para solventar esta limitación, a la hora de construir comités de redes RA-we se empleará una versión generalizada del esquema de voto por mayoría.

   Para presentar este esquema de voto generalizado, se va a formular el proceso de votación por mayoría de la siguiente manera: considérese el vector formado por las salidas de las J redes individuales para el dato $\mathbf{x}$, i.e, $\mathbf{f}(\mathbf{x}) = [f^{(1)}(\mathbf{x}), ..., f^{(J)}(\mathbf{x})]^T$, y ordénense sus componentes según sus valores (por ejemplo, situando la componente de mayor valor en la primera posición y la de menor valor en la última), obteniendo así el vector $\mathbf{f}_{\text{ord}}(\mathbf{x}) = [f^{(1)}_{\text{ord}}(\mathbf{x}), \ldots, f^{(J)}_{\text{ord}}(\mathbf{x})]^T$, donde $f^{(1)}_{\text{ord}}(\mathbf{x}) \geq f^{(2)}_{\text{ord}}(\mathbf{x}) \geq \ldots \geq f^{(J)}_{\text{ord}}(\mathbf{x})$; con el vector $\mathbf{f}_{\text{ord}}(\mathbf{x})$ se obtendrá la salida final del comité como:

$$F_{\text{voto}}(\mathbf{x}) = \sum_{j=1}^{J} w_j f^{(j)}_{\text{ord}}(\mathbf{x}) \tag{4.7}$$

   donde el vector de pesos $\mathbf{w} = [w_1, \ldots, w_J]^T$, para el caso de voto por mayoría, es un vector de ceros con un único uno en la posición $\frac{J+1}{2}$ cuando J es impar o en la posición $\frac{J}{2}$ ó $\frac{J}{2}+1$ cuando J es par. De este modo, el valor de salida será directamente el de la componente intermedia del vector $\mathbf{f}_{\text{ord}}(\mathbf{x})$, y la clase asignada aquélla en la que la mayoría de las redes están de acuerdo.

   El método de votación generalizado que se propone, al igual que el voto por mayoría, ordena las salidas de las redes individuales y calcula la salida final del comité según (4.7), pero, a diferencia de él, ajusta el vector de pesos $\mathbf{w}$ de modo que se minimice el MSE sobre S. Dado que una vez que se han ordenado las salidas de las redes

el diseño es completamente lineal, se puede emplear el mismo procedimiento de optimización que el utilizado por el método de combinación lineal, calculando el vector $\mathbf{w}$ directamente con (4.3), sin más que sustituir $\mathbf{F}$ por

$$\mathbf{F}_{\text{ord}} = \begin{bmatrix} f_{\text{ord}}^{(1)}(\mathbf{x}^{(1)}) & \dots & f_{\text{ord}}^{(J)}(\mathbf{x}^{(1)}) \\ \vdots & \ddots & \vdots \\ f_{\text{ord}}^{(1)}(\mathbf{x}^{(L)}) & \dots & f_{\text{ord}}^{(J)}(\mathbf{x}^{(L)}) \end{bmatrix} \tag{4.8}$$

4. **Construcción de comités según el criterio del RA-we**

Dado que las redes que se están combinando son conjuntos RA-we, el último criterio de combinación que se propone consiste en emplear el mismo criterio que se sigue en el diseño de estos conjuntos para la selección de los pesos de salida del comité. Como el algoritmo RA-we funciona de manera iterativa, entrenando en cada ronda un clasificador y asignándole un peso, se tendrá que proceder de manera equivalente, seleccionando, durante J rondas, una red entre el conjunto de redes disponibles, además de calcular su correspondiente peso de salida. Para seleccionar en cada ronda la red que se incorporará al comité, se ha empleado como medida de relevancia el parámetro generalizado de separación medido sobre la salida del conjunto RA-we.

Considérese que se está en la ronda $j'+1$-ésima, y que ya se han incorporado j' conjuntos RA-we al comité, concretamente, aquellas $\{f^{(j)}(\mathbf{x})\}$ con $j \in I_{j'}$, siendo $I_{j'}$ el conjunto de índices de las redes ya incluidas; quedan, por tanto, $J - j'$ conjuntos por añadir, con índices $\bar{I}_{j'}$, siendo $\bar{I}_{j'} = \{1, \dots, J\} - I_{j'}$. Para seleccionar cuál de los conjuntos restantes se añadirá en la presente iteración, se calculará para cada candidato $\{f^{(i)}(\mathbf{x})\}_{i \in \bar{I}_{j'}}$ el valor de su parámetro generalizado de separación, $\delta^{(i)}$, dado por

$$\delta^{(i)} = \frac{1}{LB^{(j')}} \sum_{l=1}^{L} \exp[-F_{\text{RA-we}}^{(j')}(\mathbf{x}^{(l)}) d^{(l)}] f^{(i)}(\mathbf{x}^{(l)}) d^{(l)} \tag{4.9}$$

donde $f^{(i)}(\mathbf{x})$ es la salida del conjunto RA-we considerado y se han utilizado

$$B^{(j')} = \frac{1}{L} \sum_{l=1}^{L} \exp[-F_{\text{RA-we}}^{(j')}(\mathbf{x}^{(l)}) d^{(l)}] \tag{4.10}$$

y

$$F^{(j')}_{\mathrm{RA-we}}(\mathbf{x}) = \sum_{j \in I_{j'}} w_j f^{(j)}(\mathbf{x}) \tag{4.11}$$

Nótese que $B^{(j')}$ y $F^{(j')}_{\mathrm{RA-we}}$ son independientes de $f^{(i)}(\mathbf{x})$. En la primera iteración del algoritmo, cuando todavía no se han incorporado redes al comité, la salida $F^{(0)}_{\mathrm{RA-we}}$ tendrá un valor nulo para cualquier dato de entrada.

A partir de los valores $\{\delta^{(i)}\}$ obtenidos, se seleccionará el conjunto RA-we que presente un valor mayor del parámetro generalizado de separación, i.e., se calculará $i^* = \mathrm{argmax}_{i \in \bar{I}_{j'}}\{\delta^{(i)}\}$ y se incorporará la red $f^{(i^*)}(\mathbf{x})$ al comité con el siguiente peso de salida

$$w^{(i^*)} = \frac{1}{2} \ln \left(\frac{1 + \delta^{(i^*)}}{1 - \delta^{(i^*)}} \right) \tag{4.12}$$

A continuación se actualizarán los conjuntos de índices $I_{j'}$ e $\bar{I}_{j'}$ y se procederá a la siguiente ronda. Una vez completadas las J rondas, la salida final del comité se obtiene como:

$$F_{\mathrm{RA-we}}(\mathbf{x}) = \sum_{j=1}^{J} w_j f^{(j)}(\mathbf{x}) \tag{4.13}$$

Las prestaciones de los comités de RA-we basados en combinación lineal y combinación lineal con función de activación tipo tangente hiperbólica ya han sido estudiados en una serie de experimentos preliminares presentados en [Gómez-Verdejo y Figueiras-Vidal, 2006], mostrando ligeras ventajas frente al RA básico y frente al CV RA-we[1]. Además, estos experimentos mostraron que algunos de los conjuntos RA-we proporcionan muy malas prestaciones y su empleo no hace sino deteriorar las del comité, por lo que sería conveniente eliminarlos de la combinación para mejorar

[1] Los métodos en [Gómez-Verdejo y Figueiras-Vidal, 2006] difieren ligeramente de los aquí presentados, ya que en ellos se empleaban las salidas normalizadas de los conjuntos RA-we; aquí se ha prescindido de esta normalización porque realmente no incorpora ventajas significativas en las dos primeras propuestas, pero sí deteriora las prestaciones del último de los métodos aquí introducidos.

los resultados finales. Para ello, en el siguiente apartado se va a presentar un método de selección de redes que, combinado con los métodos que se acaban de presentar, permite seleccionar las redes más adecuadas para formar el comité.

4.1.2. Selección de conjuntos RA-we

Tal y como se ha indicado, para formar el comité se dispone de una serie de conjuntos RA-we, cada uno de ellos entrenado con un valor diferente del parámetro de mezcla, lo que da lugar a redes con muy buenas prestaciones (las que emplean valores adecuados de λ) y otras que las tienen malas o, incluso, muy malas (cuando λ está lejos del valor óptimo). Por este motivo, resulta conveniente eliminar aquellas redes de muy malas prestaciones previamente a la construcción del comité. A pesar de que los métodos que se proponen para combinar los conjuntos RA-we tienen en cuenta las prestaciones de cada conjunto y son capaces de asignarles distinta importancia, se ha observado que la eliminación previa de estos conjuntos suele mejorar las prestaciones del comité final, además de reducir su tamaño, disminuyendo así el coste computacional.

Para analizar la influencia de cada conjunto RA-we sobre el comité final, se puede emplear la descomposición sesgo-varianza del error de generalización del comité propuesta en [Krogh y Vedelsby, 1995], que indica cómo influyen los errores de generalización de las redes que lo componen sobre el error global[2]. Concretamente, esta descomposición indica que el error de generalización del comité se puede descomponer en la diferencia de dos términos[3]: el primero de ellos, asociado al sesgo del error del comité, se corresponde con

[2] Se ha empleado la descomposición original de Krogh y Veldelsby, en lugar de la descomposición habitual (que mide el valor del MSE entre el error de generalización y el error esperado) porque esta última considera que todas las redes presentan el mismo error, mientras que la primera permite expresar el error del comité en función de los errores de las redes que lo componen, lo cual resulta de utilidad para justificar el procedimiento empleado para la selección de redes.

[3] La descomposición que se está presentando es estrictamente cierta cuando los pesos de salida del comité están normalizados (su suma es la unidad); dado que el hecho de aplicar una normalización a los

la suma ponderada de los errores de generalización de las redes que lo componen; mientras que el segundo, que se resta al anterior, está asociado a la varianza y mide la ambigüedad o diversidad existente entre las salidas de las redes. En definitiva, esta descomposición indica que resulta deseable que las redes que van a formar el comité presenten un bajo error y, además, que sus errores presenten diversidad entre sí, para que al combinarlas el error del comité pueda reducirse al máximo.

Por este motivo, la mayoría de los métodos para la construcción de comités están basados en buscar diversidad entre las redes componentes (véase [Kuncheva, 2004]); sin embargo, no se debe olvidar la importancia del término correspondiente al sesgo, el cual tiene tanta o más influencia que el término de varianza. Ante esta discrepancia, y de cara a elegir el criterio de selección de conjuntos RA-we más adecuado, se ha realizado una serie de experimentos previos para analizar la influencia que tienen sobre el comité final tanto el error (calidad) de los conjuntos RA-we, como la diversidad existente entre ellos. A partir de estos experimentos, se ha observado que resulta más adecuado emplear criterios que midan la calidad de estos conjuntos y que seleccionarlos atendiendo a la diversidad resulta, en la mayoría de los casos, poco eficaz. Por lo tanto, y a diferencia de la mayoría de métodos de construcción de comités, se ha decidido seleccionar las redes que van a formar el comité atendiendo a su calidad, la cual se medirá a través del error de entrenamiento.

De entre el conjunto de redes candidatas a formar el comité, considérese la red j-ésima, con el siguiente error de clasificación sobre el conjunto de datos de entrenamiento:

$$E^{S(j)} = \frac{1}{2L} \sum_{l=1}^{L} \mid \text{sign}[f^{(j)}(\mathbf{x}^{(l)})] - d^{(l)} \mid \tag{4.14}$$

siendo el valor de (4.14) la medida de calidad empleada en principio. Sin embargo, cuando los errores de entrenamiento de varios de los conjuntos RA-we, $\{E^{S(1)}, ..., E^{S(J)}\}$, son nulos, este parámetro de calidad presenta el inconveniente de no aportar información suficiente sobre cuál de ellas es mejor, y otras medidas de error, como el MSE, resultan

pesos de los comités de RA-we no afecta a sus prestaciones, esta descomposición permite también entender el comportamiento del error en los comités bajo estudio.

más adecuadas como parámetro de calidad[4]. Por lo tanto, cuando varios de los valores $\{E^{S^{(1)}}, ..., E^{S^{(J)}}\}$ sean nulos el parámetro de calidad utilizado será el MSE:

$$\mathrm{MSE}^{S^{(j)}} = \frac{1}{L}\sum_{l=1}^{L}\left(d^{(l)} - f^{(j)}(\mathbf{x}^{(l)})\right)^2 \tag{4.15}$$

Una vez calculado el conjunto de parámetros de calidad de las J redes RA-we (ya sea mediante (4.14) ó (4.15)), se debe establecer un umbral que permita eliminar aquellas redes con peores prestaciones y, así, formar el comité final con las restantes. Para seleccionar este umbral se ha empleado un proceso de CV (aplicado únicamente sobre la capa de salida). Por último, y dado que se desconoce el rango del conjunto de parámetros de calidad, se ha acotado el intervalo de búsqueda normalizando sus valores mediante una transformación lineal, de modo que la media muestral de los valores normalizados sea nula y su varianza muestral unitaria.

4.2. CLASIFICACIÓN ACELERADA DE COMITÉS DE CONJUNTOS RA-WE

A pesar de la reducción del coste computacional que posibilita la eliminación de las redes RA-we con malas prestaciones, los comités que se están construyendo todavía requieren un cómputo elevado a la hora de clasificar un dato, ya que cada red RA-we es, por sí misma, un conjunto de RRNN. Por este motivo, se va a presentar un método que permite paliar este inconveniente. Aunque este método se propuso en principio para redes tipo RA y redes de funciones de base radial [Arenas-García et al., 2005, Arenas-García et al., 2007], su aplicación a comités de redes RA-we puede aportar mayores ventajas que las que ya ha demostrado en otro tipo de redes.

[4]Nótese que dado que el objetivo es resolver un problema de clasificación, es más adecuado emplear (4.14) como medida de calidad y sólo emplear el MSE en aquellos casos en los que el error de clasificación no proporcione información suficiente.

Para facilitar la comprensión de dicho método, se verá, en primer lugar, en qué tipos de comités puede emplearse; a continuación, se describirá su funcionamiento y, por último, se verá cómo ordenar las redes de manera que el ahorro computacional sea lo mayor posible.

4.2.1. Compatibilidad de los comités con el método de clasificación rápida

El método de clasificación acelerada propuesto en [Arenas-García et al., 2005, Arenas-García et al., 2007] es aplicable a clasificadores binarios que calculan la clase asociada según el signo de una combinación lineal de términos, i.e., a máquinas que implementan funciones del tipo siguiente:

$$\hat{d}(\mathbf{x}) = \text{sign}\left[F(\mathbf{x})\right] = \text{sign}\left[\sum_{k=1}^{K} v_k o_k(\mathbf{x})\right] \tag{4.16}$$

En el caso de comités de redes RA-we, los términos $\{o_k(\mathbf{x})\}_{k=1}^{K}$ pueden representar las salidas de los conjuntos RA-we, pero también las salidas de cada una de las redes que componen los conjuntos RA-we, lo que permite aprovechar al máximo las ventajas de este método.

Considerando esta segunda opción, en este apartado se va a analizar, para cada uno de los métodos propuestos en el Apartado 4.1.1, cómo es la relación entre las salidas de las redes que forman los conjuntos RA-we y la etiqueta estimada por el comité, comprobando así qué tipos de comités cumplen la relación (4.16) y, por lo tanto, sobre qué tipos de comités es posible aplicar el método de clasificación acelerada.

Recuérdese que los comités bajo estudio se construyen a partir de la combinación poderada de las salidas de J conjuntos RA-we, $\{f^{(1)}(\mathbf{x}), ..., f^{(J)}(\mathbf{x})\}$, donde los valores de estas salidas se obtienen mediante la expresión

$$f^{(j)}(\mathbf{x}) = \sum_{t=1}^{T_j} \alpha_t^{(j)} o_t^{(j)}(\mathbf{x}) \tag{4.17}$$

siendo $\{o_t^{(j)}(\mathbf{x})\}_{t=1}^{T_j}$ el conjunto de los T_j clasificadores que forman el j-ésimo conjunto RA-we y $\{\alpha_t^{(j)}\}_{t=1}^{T_j}$ el conjunto de pesos asociados a dichos clasificadores.

A continuación se analiza qué tipos de comités de los presentados en el Apartado 4.1.1 dan lugar a clasificadores que pueden reescribirse según (4.16), i.e., qué comités de los allí propuestos son compatibles con el método de clasificación acelerada.

1. **Combinación lineal**

   En este primer caso, la salida del comité se puede obtener según

$$F_{\text{lin}}(\mathbf{x}) = \sum_{j=1}^{J} w_j f^{(j)}(\mathbf{x}) = \sum_{j=1}^{J} w_j \sum_{t=1}^{T_j} \alpha_t^{(j)} o_t^{(j)}(\mathbf{x}) \tag{4.18}$$

   Si se redefine el doble sumatorio como un único sumatorio con índice k (desde 1 a K, siendo $K = \sum_{j=1}^{J} T_j$), y se denota con v_k al producto $w_j \alpha_t^{(j)}$ correspondiente, se puede expresar la salida del comité como

$$F_{\text{lin}}(\mathbf{x}) = \sum_{k=1}^{K} v_k o_k(\mathbf{x}) \tag{4.19}$$

   de este modo, se incluyen en un único sumatorio las salidas de todas las redes constituyentes. Y dado que la clase estimada para $\mathbf{x}$ vendrá dada por el signo de $F_{\text{lin}}(\mathbf{x})$, se tiene que

$$\hat{d}(\mathbf{x}) = \text{sign}\left[F_{\text{lin}}(\mathbf{x})\right] = \text{sign}\left[\sum_{k=1}^{K} v_k o_k(\mathbf{x})\right] \tag{4.20}$$

   que, como se puede comprobar, es equivalente a la ecuación (4.16) y, por lo tanto, permite aplicar el método de clasificación acelerada sobre este tipo de comités.

2. **Combinación lineal con función de activación tipo tangente hiperbólica**

   Observando que la salida del comité, en este caso, viene dada por

$$F_{\text{th}}(\mathbf{x}) = \tanh\left(\sum_{j=1}^{J} w_j f^{(j)}(\mathbf{x})\right) \tag{4.21}$$

se puede ver que este procedimiento de combinación es equivalente a (4.18) y (4.19), salvo por la aplicación de la función tangente hiperbólica. Dado que dicha función de activación únicamente influye en el procedimiento de determinación de los $\{w_j\}$, pero no tiene influencia alguna en el criterio de clasificación implementado por la red, es inmediato concluir que este tipo de comités también resulta compatible con el método de clasificación rápida.

3. **Esquema de voto generalizado**

   La salida del comité viene dada por

$$F_{\text{voto}}(\mathbf{x}) = \sum_{j=1}^{J} w_j f_{\text{ord}}^{(j)}(\mathbf{x}) \tag{4.22}$$

   donde el conjunto de valores $\{f_{\text{ord}}^{(j)}(\mathbf{x})\}_{j=1}^{J}$ se corresponde con el conjunto de las salidas de las J redes RA-we, $\{f^{(j)}(\mathbf{x})\}_{j=1}^{J}$, cuando estas han sido ordenadas según sus valores. El proceso de reordenamiento requiere la evaluación de todas las redes, y es un proceso no lineal (la salida no es una combinación lineal fija de las salidas de las redes), lo que impide aplicar el método de clasificación acelerada en este tipo de comité.

4. **Criterio del RA-we**

   En este último caso la salida del comité se obtiene de forma equivalente al caso de combinación lineal, i.e.,

$$F_{\text{RA}-\text{we}}(\mathbf{x}) = \sum_{j=1}^{J} w_j f^{(j)}(\mathbf{x}) \tag{4.23}$$

   Por este motivo, se puede aplicar el mismo razonamiento seguido para el método de combinación lineal, el método de clasificación acelerada resulta compatible con este tipo de comités.

4.2.2. Procedimiento de clasificación acelerada

Teniendo en cuenta que la salida de los diferentes tipos de comités, excepto para el esquema de voto generalizado, se puede expresar mediante la combinación lineal de K términos, el método que se propone permite acelerar el cálculo de (4.16) mediante una evaluación secuencial de los diferentes términos que componen el sumatorio hasta que se disponga de una confianza suficiente sobre la clase a la que pertenece el patrón, momento en el que se sigue el criterio proporcionado por la salida parcial del sistema

$$F_k(\mathbf{x}) = \sum_{k'=1}^{k} v_{k'} o_{k'}(\mathbf{x}) \tag{4.24}$$

para predecir la clase a la que pertenece $\mathbf{x}$, en vez de utilizar la salida global, $F(\mathbf{x})$. De este modo, no es necesario evaluar los $K - k$ términos restantes del sumatorio y, consecuentemente, se reduce el coste computacional.

Sin embargo, el hecho de emplear la salida parcial del sistema en lugar de utilizar la salida global causa el siguiente error sobre el valor de salida obtenido para $\mathbf{x}$:

$$e_k(\mathbf{x}) = F_K(\mathbf{x}) - F_k(\mathbf{x}) = \sum_{k'=k+1}^{K} v_{k'} o_{k'}(\mathbf{x}) \tag{4.25}$$

cuyo valor puede acotarse, teniendo en cuenta que las salidas de las redes RA-we se encuentran dentro del intervalo $[-1, 1]$, de manera que:

$$|e_k(\mathbf{x})| \leq \sum_{k'=k+1}^{K} |v_{k'}| = \eta_k \tag{4.26}$$

Esta cota indica que cuando la salida parcial de sistema $F_k(\mathbf{x})$ tiene un valor absoluto mayor que el umbral η_k, su signo no va a cambiar aunque se evalúen los $K - k$ términos restantes y, por lo tanto, la etiqueta estimada por la salida parcial va a coincidir con la obtenida por el comité completo. Sin embargo, si $|F_k(\mathbf{x})| < \eta_k$ se debe proceder a evaluar los siguientes términos del sumatorio hasta que se cumpla la condición

$$|F_k(\mathbf{x})| > \eta_k \tag{4.27}$$

En la mayoría de los problemas de clasificación existen datos fáciles de clasificar, para los cuales sólo será necesario evaluar los primeros términos del sumatorio, obteniendo para estos casos una gran reducción computacional; para otros datos, será necesaria la evaluación de todos o casi todos los términos del sumatorio y la reducción computacional será menor. En cualquier caso, el procedimiento descrito garantiza que no se modifica el criterio original de la red.

Si se quisiera reducir aún más el coste computacional, se podría disminuir el valor del umbral, suavizando la condición (4.27) con la inclusión de un parámetro $\beta \in [0, 1]$. De esta manera, se tomará la condición

$$|F_k(\mathbf{x})| > \beta\eta_k \tag{4.28}$$

en vez de la condición (4.27). Obviamente, este criterio puede modificar la función de clasificación del comité completo; no obstante, obsérvese que la cota (4.26) es muy conservadora ya que asume que $o_{k'}(\mathbf{x}) = \pm 1, \quad k' = k+1, \ldots, K$, por lo que puede esperarse que la degradación introducida no sea muy importante. Lógicamente la selección de β supone un compromiso entre le ahorro computacional obtenido (mayor según decrece β) y el respeto al criterio del comité original (que únicamente se garantiza para $\beta = 1$).

En el Cuadro 4.1 se recoge el pseudocódigo del algoritmo que se acaba de presentar.

4.2.3. Reordenamiento de las redes para máximo ahorro computacional

Si se desea explotar al máximo las ventajas que el método de clasificación acelerada puede aportar, es necesario ordenar las redes que forman el sistema adecuadamente, de modo que las primeras redes en ser evaluadas sean las más relevantes (las que tengan una mayor importancia sobre la salida global del sistema), consiguiendo así que el número de redes que se tenga que evaluar sea, en promedio, el menor posible.

En este caso, es necesario emplear una medida o criterio de relevancia que nos indique cómo se deben ordenar las redes. Entre los múltiples criterios que se podrían

Cuadro 4.1: Pseudocódigo del procedimiento de clasificación acelerada en comités.

1 - Entradas:

· Nuevo dato de entrada, $\mathbf{x}$.

· Sistema de RRNN ya entrenado formado por el conjunto de redes $\{o_k(\mathbf{x})\}_{k=1}^{K}$ y el conjunto de pesos $\{v_k(\mathbf{x})\}_{k=1}^{K}$.

· Parámetro suavizador del umbral β.

2 - Calcular los valores $\{\eta_k\}_{k=1}^{K}$

3 - Para $k = 1, \ldots, K$

3.1 - Obtener la salida de la red k-ésima para el dato $\mathbf{x}$: $o_k(\mathbf{x})$.

3.1 - Calcular la salida parcial del sistema:

$$F_k(\mathbf{x}) = \sum_{k'=1}^{k} v_{k'} o_{k'}(\mathbf{x})$$

3.3 - Si $|F_k(\mathbf{x})| \geq \beta\eta_k$,

detener el proceso y establecer:

$$F_K(\mathbf{x}) = F_k(\mathbf{x})$$

3 - Predecir la clase a la que pertenece $\mathbf{x}$ como:

$$\hat{d}(\mathbf{x}) = \text{sign}\left[F_K(\mathbf{x})\right]$$

emplear, medidas de correlación o medidas sobre el valor medio de las salidas (véase [Arenas-García et al., 2007], donde se describen varias de estas alternativas), la opción que se va a emplear consiste en ordenar las redes por orden decreciente de los valores $|v_k|$ que tienen asociados. Este método no sólo resulta de gran sencillez sino que, además, tiene una base teórica, ya que equivale a considerar la influencia que la salida de cada una de las redes tiene sobre la función de coste empleada para el diseño del comité.

Para mostrar esta equivalencia, puede calcularse, en primer lugar, la influencia de cada red $o_k(\mathbf{x})$ sobre una función de coste genérica C, derivando dicha función respecto a la

salida de la red k-ésima para el dato l-ésimo de S, es decir, calculando $\partial C/\partial o_k(\mathbf{x}^{(l)})$. Para el cálculo de esta derivada, considérese que la salida del comité para el dato l-ésimo, independientemente del tipo de comité, viene dada por $F(\mathbf{x}^{(l)})$, de manera que

$$\frac{\partial C}{\partial o_k(\mathbf{x}^{(l)})} = \frac{\partial C}{\partial F(\mathbf{x}^{(l)})}\frac{\partial F(\mathbf{x}^{(l)})}{\partial o_k(\mathbf{x}^{(l)})} = \frac{\partial C}{\partial F(\mathbf{x}^{(l)})}F'(\mathbf{x}^{(l)})v_k \tag{4.29}$$

donde $F'(\mathbf{x}^{(l)})$ denota la derivada de la función de activación que tenga el comité (en los casos lineales tendrá valor unitario, y en el caso de emplear la tangente hiperbólica su valor será $\left[1 - F_{\text{th}}^2(\mathbf{x}^{(l)})\right]$).

Para obtener un valor numérico que indique la influencia de $o_k(\mathbf{x})$ sobre la función de coste, se puede construir el gradiente formado por estas derivadas parciales, i.e,

$$\nabla_{o_k} C = v_k \left[\frac{\partial C}{\partial F(\mathbf{x}^{(1)})}F'(\mathbf{x}^{(1)}), \ldots, \frac{\partial C}{\partial F(\mathbf{x}^{(L)})}F'(\mathbf{x}^{(L)})\right] \tag{4.30}$$

y calcular su norma

$$\|\nabla_{o_k} C\| = |v_k| \sqrt{\sum_{l=1}^{L} \frac{\partial C}{\partial F(\mathbf{x}^{(l)})}F'_{act}(\mathbf{x}^{(l)})} \tag{4.31}$$

Teniendo en cuenta que el valor de la raíz cuadrada es independiente de la red analizada, la relevancia de $o_k(\mathbf{x})$ respecto al resto de redes dependerá únicamente del valor absoluto del peso v_k que tenga asociado, justificando el criterio seleccionado para el reordenamiento de las unidades.

4.3. CONCLUSIONES

En este capítulo se ha presentado una serie de métodos para construir comités formados por conjuntos RA-we, con lo que se evita el problema de seleccionar adecuadamente el parámetro de mezcla λ y, además, se aprovecha la diversidad existente entre conjuntos RA-we entrenados con distintos valores de λ.

Concretamente, se han planteado cuatro métodos que permiten realizar la agrupación de conjuntos RA-we utilizando, cada uno de ellos, un criterio distinto para obtener los

pesos de salida del comité. Además, junto con estos métodos, se ha propuesto aplicar un método para selección de los conjuntos RA-we que deben formar el comité, eliminándose aquéllos que podrían deteriorar las prestaciones finales del comité.

Para finalizar, se ha considerado un método que permite realizar una clasificación acelerada de los datos, de gran utilidad para reducir el elevado coste computacional que estos comités pueden presentar durante su fase operacional.

Capítulo 5

Ajuste dinámico de la función de énfasis

En este capítulo se considerará una alternativa diferente para la selección del parámetro de mezcla de la función de énfasis mixto, que, como ya se ha discutido, es un problema fundamental para obtener buenas prestaciones de un esquema RA-we, y no se resuelve de modo completamente satisfactorio mediante CV. Se partirá de los fundamentos teóricos que justifican el buen comportamiento del RA-we para proponer una alternativa completamente diferente al uso de CV o la construcción de comités. Concretamente, se propone realizar una selección dinámica del parámetro de mezcla, i.e., eligiendo en cada iteración el valor que proporciona las mejores prestaciones al conjunto construido hasta el momento. Se denominará al diseño resultante DW-RA ("Dynamically adapted Weighted emphasis version of Real Adaboost").

En la última sección del capítulo se analizarán las ventajas que la selección dinámica aporta frente al RA-we, que emplea un valor fijo del parámetro de mezcla y que incluye como caso particular al RA.

5.1. EL ALGORITMO DW-RA

El algoritmo RA-we propuesto en el Capítulo 3 requiere el cálculo en cada iteración de un parámetro δ_t, denominado parámetro generalizado de separación del clasificador, que viene dado por la expresión:

$$\delta_t = \frac{1}{LB_{t-1}} \sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]o_t(\mathbf{x}^{(l)})d^{(l)} \tag{5.1}$$

Según se explicó, este parámetro mide la calidad del clasificador t-ésimo, ya que calcula la correlación existente entre las salidas del clasificador para el conjunto de datos de entrenamiento y sus correspondientes etiquetas (ponderada por la contribución parcial de cada dato a B_{t-1}). Por este motivo, se va a emplear δ_t como parámetro de calidad de los clasificadores, seleccionando en cada iteración el valor del parámetro de mezcla, λ, que proporciona el valor más elevado de δ_t.

Concretamente, el algoritmo que se propone opera del siguiente modo: en cada ronda se entrena un conjunto de clasificadores, $\{o_t^{(j)}(\mathbf{x})\}_{j=1}^{J}$, cada uno de ellos asociado a un valor del parámetro de mezcla que es seleccionado entre un conjunto predefinido de valores, $\{\lambda^{(j)}\}_{j=1}^{J}$. De este modo, cada clasificador aprende empleando la función de énfasis mixto (3.6) particularizada para el valor $\lambda^{(j)}$ que tenga asociado, i.e.,

$$D_{\lambda,t+1}^{(j)}(\mathbf{x}^{(l)}) = \frac{1}{Z_{\lambda,t}^{(j)}} \exp\left\{\lambda^{(j)}[f_t(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1-\lambda^{(j)})f_t^2(\mathbf{x}^{(l)})\right\} \tag{5.2}$$

Una vez entrenados estos clasificadores, se calcula el valor de δ_t asociado a cada clasificador, obteniendo así un conjunto de valores de parámetros de separación generalizados, $\{\delta_t^{(j)}\}_{j=1}^{J}$, y, consecuentemente, un conjunto de posibles pesos de salida, $\{\alpha_t^{(j)}\}_{j=1}^{J}$, donde cada uno de estos valores se determina siguiendo el mismo criterio empleado por el RA-we pero particularizado para el correspondiente $\delta_t^{(j)}$; es decir,

$$\alpha_t^{(j)} = \frac{1}{2}\ln\left(\frac{1+\delta_t^{(j)}}{1-\delta_t^{(j)}}\right) \tag{5.3}$$

El conjunto de valores $\{\delta_t^{(j)}\}_{j=1}^J$ permite seleccionar el clasificador que ha de incorporarse al conjunto como aquél que resulta en un valor mayor del parámetro generalizado de separación, i.e., el de mayor calidad. Dado que $\delta_t \in [0, 1)$ y (5.3) es una función monótona creciente de $\delta_t^{(j)}$ en dicho intervalo, un criterio equivalente consiste en seleccionar el clasificador que tiene asociado un mayor peso de salida y que, por tanto, influirá de forma más significativa en el conjunto. Por tanto, si se define

$$j^* = \arg \max_j \ \delta_t^{(j)} = \arg \max_j \ \alpha_t^{(j)} \tag{5.4}$$

el algoritmo DW-RA añadirá al conjunto el clasificador $o_t(\mathbf{x}) = o_t^{(j^*)}(\mathbf{x})$ con un peso $\alpha_t = \alpha_t^{(j^*)}$. El parámetro de mezcla y de separación generalizado asociados a dicho clasificador pueden denotarse, respectivamente, como $\lambda_t = \lambda^{(j^*)}$ y $\delta_t = \delta_t^{(j^*)}$.

Nótese que este algoritmo, en comparación con el RA-we, no sólo proporciona un criterio automático de ajuste de la función de énfasis, sino que permite modificar el valor del parámetro de mezcla empleado en cada iteración.

En el Cuadro 5.1 se resumen los pasos de este algoritmo al que se ha denominado RA-we con selección dinámica del parámetro de mezcla ("Dynamically adapted Weighted emphasis version of Real Adaboost") y que se denotará como DW-RA.

Por último, debe indicarse que la selección del mejor parámetro de mezcla en cada iteración no tiene por qué aportar el mejor diseño desde un punto de vista global; sin embargo, tal y como se ha comprobado experimentalmente, este método proporciona mejores resultados que otros criterios para la selección dinámica de λ, como, por ejemplo, la selección del clasificador asociado a la mediana del conjunto $\{\delta_t^{(j)}\}_{j=1}^J$. Además de la justificación empírica con un elevado número de experimentos que se expondrán en el Capítulo 6, la idoneidad del criterio seleccionado encuentra explicación en una serie de resultados teóricos que se discuten en la sección siguiente.

Cuadro 5.1: Pseudocódigo de funcionamiento del algoritmo DW-RA.

1 - Entradas: $\{\mathbf{x}^{(l)}, d^{(l)}\}_{l=1}^{L}$, $\{\lambda^{(j)}\}_{j=1}^{J}$

2 - Entrenar el primer clasificador, $o_1(\mathbf{x})$, minimizando la función de coste:

$$C_1 = \frac{1}{L}\sum_{l=1}^{L}[d^{(l)} - o_1(\mathbf{x}^{(l)})]^2$$

y añadirlo al conjunto calculando: $\alpha_1 = \frac{1}{2}\ln\left(\frac{1+\delta_1}{1-\delta_1}\right)$, donde $\delta_1 = \frac{1}{L}\sum_{l=1}^{L} o_t^{(j)}(\mathbf{x}^{(l)})d^{(l)}$.

3 - Para $t = 2, \ldots, T$

3.1 - Para $j = 1, \ldots, J$

3.1.1 - Calcular la función de énfasis correspondiente a $\lambda^{(j)}$:

$$D_{\lambda,t}^{(j)}(\mathbf{x}^{(l)}) = \frac{1}{Z_{\lambda,t}^{(j)}}\exp\left\{\lambda^{(j)}[f_{t-1}(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1-\lambda^{(j)})f_{t-1}^2(\mathbf{x}^{(l)})\right\}$$

3.1.2 - Entrenar $o_t^{(j)}(\mathbf{x}^{(l)})$ minimizando la función de coste:

$$C_t^{(j)} = \sum_{l=1}^{L} D_{\lambda,t}^{(j)}(\mathbf{x}^{(l)})[d^{(l)} - o_t^{(j)}(\mathbf{x}^{(l)})]^2$$

3.1.3 - Calcular el parámetro generalizado de separación asociado a $o_t^{(j)}(\mathbf{x}^{(l)})$

$$\delta_t^{(j)} = \frac{1}{LB_{t-1}}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]o_t^{(j)}(\mathbf{x}^{(l)})d^{(l)}$$

donde $B_{t-1} = \frac{1}{L}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]$.

3.1.4 - Calcular el peso de salida asociado al clasificador

$$\alpha_t^{(j)} = \frac{1}{2}\ln\left(\frac{1+\delta_t^{(j)}}{1-\delta_t^{(j)}}\right)$$

3.2 - Añadir al conjunto:

$$o_t(\mathbf{x}) = o_t^{(j^*)}(\mathbf{x}) \text{ con } \alpha_t = \alpha_t^{(j^*)}$$

donde $j^* = \arg\max_j \delta_t^{(j)}$

4 - El clasificador final implementa la función

$$f_T(\mathbf{x}) = \sum_{t=1}^{T}\alpha_t o_t(\mathbf{x})$$

siendo, por tanto, $\hat{d}(\mathbf{x}) = \text{sign}\,[f_T(\mathbf{x})]$ la clase estimada para el patrón $\mathbf{x}$.

5.2. VENTAJAS DE DW-RA FRENTE A RA-WE

En el Capítulo 2 se revisaron dos resultados teóricos que justifican el buen comportamiento del algoritmo RA, referentes a la convergencia del error de entrenamiento y al comportamiento del error de generalización. En el Capítulo 3 se vio cómo estas dos propiedades podían extenderse al algoritmo RA-we de forma más o menos directa.

A continuación, se mostrará cómo estas propiedades siguen verificándose para el algoritmo que se acaba de proponer, mostrando, además, cómo la particular manera que tiene el DW-RA de seleccionar el valor de λ permite esperar mejores resultados desde un punto de vista de velocidad de convergencia en el error de entrenamiento, así como en la capacidad de generalización, en comparación con el RA-we y el RA.

5.2.1. Aceleración de la reducción del error de entrenamiento

Tal y como se vio en el Apartado 3.4.1, y debido al criterio empleado para la selección de los valores $\{\alpha_t\}$, para el caso del algoritmo RA-we se puede acotar el error de entrenamiento E_t^S mediante:

$$E_t^S = \frac{1}{2L}\sum_{l=1}^{L} \mid \text{sign}\left[f_t(\mathbf{x}^{(l)})\right] - d^{(l)} \mid \leq B_t \leq \prod_{t'=1}^{t}\sqrt{1-\delta_{t'}^2} \tag{5.5}$$

La demostración de esta cota, presentada en el Anexo A.2, se basa en la aplicación recursiva de la cota: $B_t \leq \sqrt{1-\delta_t^2} \quad B_{t-1}$, cuya validez depende únicamente del criterio empleado para la selección de α_t: dicha cota es válida siempre que los pesos se calculen de acuerdo a la expresión (3.9). Dado que el algoritmo DW-RA calcula sus pesos utilizando esa misma expresión, independientemente del valor λ_t seleccionado, la cota (5.5) también puede aplicarse al error de entrenamiento en que incurre el algoritmo DW-RA.

Si se analiza el valor de (5.5) para los distintos algoritmos considerados, es inmediato comprobar que, dado que el DW-RA selecciona en cada iteración el valor máximo posible para el parámetro generalizado de separación, es de esperar que los valores de $\{\delta_t\}$ sean mayores para este algoritmo que para el RA-we (y que los valores $\{\gamma_t\}$ para el algoritmo

RA), lo que resulta en un menor valor de la cota. En otras palabras, se puede esperar que DW-RA presente una mayor velocidad en cuanto a la reducción del error de entrenamiento que los algoritmos analizados previamente.

Es interesante observar que la minimización de la cota sobre el error de entrenamiento junto con la aplicación del método de Gauss-Southwell [Luenberger, 1984] proporcionan una justificación adicional del criterio seguido para la selección de λ_t. El método de Gauss-Southwell se utiliza para minimizar una función de J variables, $g(x_1, ..., x_J)$ de forma iterativa mediante la optimización en cada paso de la variable que presenta un mayor valor absoluto del gradiente de $g(x_1, ..., x_J)$, seleccionando, para cada iteración, la componente que permite una mejor aproximación al mínimo de $g(x_1, ..., x_J)$. De manera formal, si se define

$$j^* = \arg\max_j \left| \frac{\partial g(x_1, ..., x_J)}{\partial x_j} \right| \tag{5.6}$$

donde $|\cdot|$ representa el operador valor absoluto, el método optimiza el valor de la variable x_{j^*}, obteniendo así un nuevo valor de la función $g(x_1, ..., x_{j^*} + \Delta x_{j^*}, ..., x_J)$, el cual será el punto de partida para la siguiente iteración.

El método de Gauss-Southwell puede aplicarse para justificar la selección del clasificador que provoca una mayor reducción de la cota de entrenamiento. Para ello, supóngase que se está en la iteración t-ésima y que se dispone de J clasificadores entrenados para ser añadidos al conjunto junto con unos pesos a determinar; en este caso, la cota sobre el error de entrenamiento del conjunto resultante de añadir estos J clasificadores es:

$$B_t^{(1,...,J)} = \frac{1}{L} \sum_{l=1}^{L} \exp\left[-f_{t-1}(\mathbf{x}^{(l)}) d^{(l)}\right] \exp\left[-\sum_{j=1}^{J} \alpha_t^{(j)} o_t^{(j)}(\mathbf{x}^{(l)}) d^{(l)}\right] \tag{5.7}$$

Considerando esta cota una función de los pesos de salida de los J clasificadores, i.e., $g(\alpha_1, ..., \alpha_J)$, la aplicación sobre ella de una iteración del método de Gauss-Southwell indicará cual de los J clasificadores consigue una mayor reducción de la cota. Para ello, pártase de la solución trivial $\alpha_t^{(j)} = 0 \; \forall j$, correspondiente al clasificador construido en la iteración $t-1$, y evalúense las derivadas de g respecto a sus parámetros:

$$\begin{aligned}
\left.\frac{\partial g}{\partial \alpha_t^{(j)}}\right|_{\alpha_t^{(j)}=0} &= \frac{1}{L}\sum_{l=1}^{L}\exp\left[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}\right]\left.\frac{\partial \exp\left[-\sum_{j=1}^{J}\alpha_t^{(j)}o_t^{(j)}(\mathbf{x}^{(l)})d^{(l)}\right]}{\partial \alpha_t^{(j)}}\right|_{\alpha_t^{(j)}=0} \\
&= -\frac{1}{L}\sum_{l=1}^{L}\exp\left[-f_t^{(j)}(\mathbf{x}^{(l)})d^{(l)}\right]o_t^{(j)}(\mathbf{x}^{(l)})d^{(l)}\Bigg|_{\alpha_t^{(j)}=0} \\
&= -\frac{1}{L}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]o_t^{(j)}(\mathbf{x}^{(l)})d^{(l)} \\
&= -B_{t-1}\delta_t^{(j)} \qquad (5.8)
\end{aligned}$$

Dado que tanto B_{t-1} como $\delta_t^{(j)}$ toman valores positivos[1], la variable $\alpha_t^{(j)}$ elegida por el método de Gauss-Southwell será aquella asociada a un mayor parámetro generalizado de separación , coincidiendo con el criterio de selección utilizado por el DW-RA. En otras palabras, el procedimiento de selección de DW-RA puede interpretarse como la aplicación, en cada etapa del crecimiento del conjunto, de una iteración del algoritmo de Gauss-Southwell, en la que se añade el aprendedor base que proporciona una menor cota sobre el error de entrenamiento. Además, según se sabe, el valor seleccionado para el correspondiente peso es precisamente el que minimiza el valor de dicha cota.

En el caso de los algoritmos RA-we y RA, por el contrario, la selección del énfasis para el nuevo aprendedor está predeterminada, y no responde a un criterio de optimización de la cota del error de entrenamiento. Cabe esperar, por tanto, que la reducción de dicha cota sea más lenta que para el algoritmo DW-RA.

[1]La definición de δ_t no garantiza que un clasificador vaya a presentar siempre valores positivos del parámetro generalizado de separación. Sin embargo, experimentalmente, se ha observado que su valor es positivo durante el crecimiento de la red (cuando el error de entrenamiento está decreciendo) y que cuando el algoritmo ha convergido (el error de entrenamiento se estanca) algunos clasificadores base pueden tener muy malas prestaciones, presentando valores de δ_t muy cercanos a cero, normalmente positivos pero, en raras ocasiones, negativos; dado que son casos muy particulares, y únicamente aparecen cuando el algoritmo ha convergido, se puede considerar $\delta_t > 0$ durante la convergencia del error de entrenamiento.

5.2.2. Mejora en la capacidad de generalización

En el Apartado 3.4.2 se vio cómo el error de generalización de un conjunto tipo RA-we, al igual que ocurría en los conjuntos tipo RA, podía analizarse en términos del riesgo marginal $R_T^{\text{margin}}(\theta)$ en vez de en términos del riesgo esperado $R[f_T]$, lo que permitía concluir que una reducción en dicho riesgo marginal posibilita la mejora de la capacidad de generalización del conjunto. De nuevo, este resultado es aplicable al algoritmo DW-RA (ya que la salida global del conjunto sigue siendo una combinación lineal de los clasificadores que lo componen). De hecho, la particular selección que realiza el DW-RA del parámetro de mezcla permite conjeturar que la capacidad de generalización en este tipo de conjuntos es generalmente mejor que la que presentan los conjuntos construidos empleando el algoritmo RA-we. Para analizar la influencia del criterio empleado por el DW-RA para la selección de λ sobre el riesgo marginal, considérese la cota establecida sobre dicho riesgo para el algoritmo RA-we y empléese sobre este nuevo algoritmo,

$$R_T^{\text{margin}}(\theta) \leq \prod_{t=1}^{T} (1+\delta_t)^{\frac{1+\theta}{2}} (1-\delta_t)^{\frac{1-\theta}{2}} \tag{5.9}$$

Esta desigualdad es directamente aplicable al DW-RA, dado que para su demostración (véase Anexo A.3) se emplea, por un lado, la desigualdad (3.11) (la cual es válida también para el DW-RA, tal y como se acaba de ver en el apartado anterior) y, por el otro, la expresión que permite calcular α_t (que para el algoritmo DW-RA es idéntica a la del RA-we).

Tal y como se vio para el algoritmo RA-we y para el RA, a partir de este resultado, se puede afirmar que $R_T^{\text{margin}}(\theta)$ tiende a decrecer con el número de rondas, cuando los factores que lo componen son menores que 1, siendo más plausible que tal cosa ocurra cuando se manejan valores pequeños de θ, y dependiendo, en última instancia, del valor del parámetro generalizado de separación. Por este motivo, se va a analizar el comportamiento del término t-ésimo de la cota (5.9) en función de δ_t. Para ello, denótese este término como $h(\delta_t)$

$$h(\delta_t) = (1+\delta_t)^{\frac{1+\theta}{2}} (1-\delta_t)^{\frac{1-\theta}{2}} \tag{5.10}$$

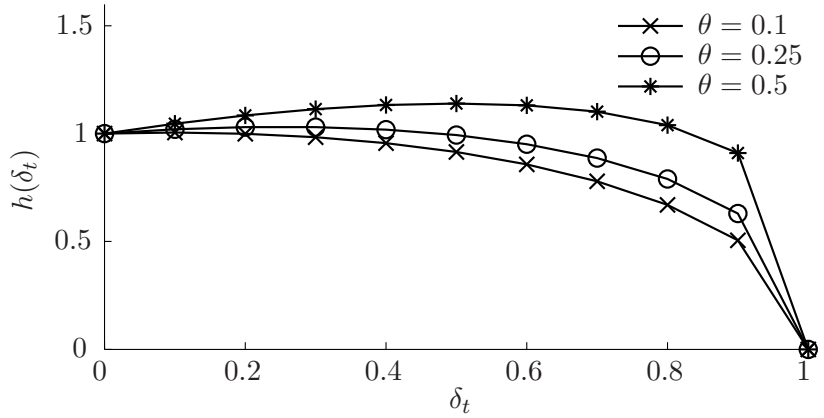

Figura 5.1: Evolución de $h(\delta_t)$ en función del parámetro generalizado de separación.

y analícese su crecimiento. Si se deriva, en primer lugar, el logaritmo neperiano de $h(\delta_t)$ y se iguala a 0 el resultado, se obtiene la siguiente ecuación:

$$\frac{\partial \ln\left[h(\delta_t)\right]}{\partial \delta_t} = \frac{1+\theta}{2}\frac{1}{1+\delta_t} - \frac{1-\theta}{2}\frac{1}{1-\delta_t} = 0 \tag{5.11}$$

que tiene por solución el punto $\delta_t = \theta$, lo que permite afirmar que en ese punto hay un extremo relativo. Examinando el valor de la derivada segunda de $h(\delta_t)$ en $\delta_t = \theta$

$$\begin{aligned} \left.\frac{\partial^2 \ln\left[h(\delta_t)\right]}{\partial^2 \delta_t}\right|_{\delta_t=\theta} &= -\frac{1+\theta}{2}\frac{1}{(1+\delta_t)^2} - \frac{1-\theta}{2}\frac{1}{(1-\delta_t)^2}\Bigg|_{\delta_t=\theta} \\ &= -\frac{1}{2(1+\theta)} - \frac{1}{2(1-\theta)} = -\frac{1}{1-\theta^2} \end{aligned} \tag{5.12}$$

y recordando además que $\theta \in (0, 1]$, se comprueba que la derivada segunda es negativa y, por lo tanto, se puede afirmar que el punto $\delta_t = \theta$ se corresponde con un máximo de $h(\delta_t)$.

En la Figura 5.1 se ilustra el comportamiento de $h(\delta_t)$ para distintos valores de θ, evidenciándose que cuando $\delta_t > \theta$ el término t-ésimo de (5.9), $h(\delta_t)$, es decreciente con δ_t, presentando un descenso bastante rápido cuando los valores de δ_t son próximos a 1, mientras que para $\delta_t < \theta$ $h(\delta_t)$ se mantiene ligeramente por encima de 1. Este resultado indica que, en cada iteración, interesa que los valores de δ_t sean lo más elevados posible para conseguir menores valores de la cota sobre el riesgo marginal.

A la luz de este análisis hay que decir que, dado que el criterio empleado por el DW-RA para la elección de λ consiste en elegir el clasificador con mayor parámetro generalizado

de separación, es esperable que dicho algoritmo sea más eficiente que los algoritmos RA-we y RA a la hora de minimizar el riesgo marginal, presentando así un menor error de generalización.

5.3. CONCLUSIONES

En este capítulo se ha propuesto un nuevo algoritmo denominado DW-RA que solventa el problema de la selección del parámetro de mezcla, λ, inherente al RA-we. Esta propuesta consiste en una selección dinámica del parámetro de la función de énfasis mixto, eligiendo en cada iteración el valor que proporcione el clasificador con mejores prestaciones (medido en términos del parámetro generalizado de separación de cada posible clasificador).

También se ha realizado un análisis teórico de las ventajas que el DW-RA puede presentar frente al RA-we y frente al RA básico, mostrando que es razonable esperar que el DW-RA presente mayor velocidad de convergencia del error de entrenamiento, mejorando además la capacidad de generalización del conjunto.

CAPÍTULO 6

EVALUACIÓN DE LAS DISTINTAS PROPUESTAS

En este capítulo, a través de una serie de experimentos, se evaluarán las prestaciones de los distintos algoritmos que se han propuesto a lo largo de esta Tesis Doctoral. En primer lugar, se describirán cómo se han llevado a cabo los experimentos: las bases de datos empleadas, el proceso de entrenamiento de los diferentes conjuntos (RA, RA-we y DW-RA) y el método para medir las diferencias estadísticas entre las tasas de error proporcionadas por cada algoritmo. Tras ello, se comparará el algoritmo RA-we frente al RA básico, verificando su ventaja potencial y analizando la efectividad de la validación cruzada a la hora de seleccionar el valor del parámetro de mezcla, λ. A continuación, se considerarán los diferentes métodos de construcción de comités que se propusieron en el Capítulo 4, lo que permitirá apreciar las ventajas que el método de clasificación acelerada puede aportar a los comités de conjuntos RA-we. Después, se procederá con el algoritmo DW-RA, encontrando que no sólo proporciona reducción de la tasa de error, sino también otras ventajas en términos de aceleración de la convergencia y de mejora en

la capacidad de generalización respecto al RA y el RA-we (para un valor fijo de λ). Por último, se presentarán las conclusiones de estos experimentos, discutiendo conjuntamente las ventajas y limitaciones de las dos propuestas principales de esta Tesis Doctoral: los comités de conjuntos RA-we y el algoritmo DW-RA.

6.1. DESCRIPCIÓN DE LOS EXPERIMENTOS

6.1.1. Bases de datos empleadas

Las prestaciones de los algoritmos se van a evaluar sobre un conjunto de 8 bases de datos binarias: cinco bases de datos pertenecientes al repositorio de la Universidad de California en Irvine (UCI) [Newman et al., 1998]: *Abalone*, *Contraceptive*, *Image*, *Spam* y *Tictactoe*; el problema *Phoneme* obtenido de [Alinat, 1993]; y dos problemas sintéticos: *Kwok* [Kwok, 1999] y *Ripley* [Ripley, 1994].

En el Cuadro 6.1 se resumen las principales características de cada uno de estos problemas: la notación o abreviatura con la que se va a denotar cada problema de ahora en adelante, su número de dimensiones (*dim*) y el número de datos que hay en cada clase, tanto en el conjunto de entrenamiento como en el de test; además, para cada uno de estos problemas se proporciona la tasa de error alcanzada por los diseños óptimos tipo Máquina de Vectores Soporte (SVM) para cada caso. Las SVM son una tecnología aceptada como el "estado del arte" en aprendizaje máquina, por lo que su tasa de error servirá de indicador para apreciar la calidad de los resultados obtenidos por los algoritmos propuestos. En el Apéndice B se detalla en qué consiste cada uno de estos problemas.

Para el entrenamiento de las SVM se ha utilizado el software diseñado y escrito por F. Pérez-Cruz [Pérez-Cruz, 2002], basado en el algoritmo "Iterative Re-Weighted Least Squares" (IRWLS) descrito en [Pérez-Cruz et al., 2001]. Además, se ha elegido un núcleo gaussiano y los parámetros de las redes (dispersión del núcleo y factor de penalización del riesgo empírico) se han seleccionado mediante validación cruzada (CV).

Cuadro 6.1: Características de las bases de datos empleadas en la evaluación de las técnicas propuestas.

Problema	Notación	*dim*	Datos de entrenamiento (n_1/n_{-1})	Datos de test (n_1/n_{-1})	Tasa de error de SVM
Abalone	*Ab*	8	1238/1269	843/827	20.9
Contraceptive	*Co*	9	506/377	338/252	28.61
Image	*Im*	18	821/1027	169/293	3.47
Kwok	*Kw*	2	300/200	6120/4080	11.74
Phoneme	*Ph*	5	952/2291	634/1527	15.35
Ripley	*Ri*	2	125/125	500/500	9.8
Spam	*Sp*	57	1673/1088	1115/725	7.2
Tictactoe	*Ti*	9	199/376	133/250	1.7

6.1.2. Entrenamiento de los conjuntos RA, RA-we y DW-RA

Para el diseño de los conjuntos RA, RA-we y DW-RA, se emplean como clasificadores base redes neuronales de tipo Perceptrón Multi Capa ("Multi Layer Perceptron", MLP). En el siguiente subapartado, se indican los parámetros de diseño y de entrenamiento empleados para llevar a cabo estos experimentos.

El entrenamiento de los conjuntos se realiza según indica el pseudocódigo recogido en los Cuadros 2.1, 3.1 ó 5.1, según se trate del algoritmo RA, RA-we o DW-RA, respectivamente, en los que se supone conocido el número de clasificadores base que constituyen el conjunto, T. En la segunda subsección se describe cómo se lleva a cabo esta selección durante el entrenamiento de las redes.

Diseño de los clasificadores base

La arquitectura de los MLPs empleados consiste en una única capa oculta, cuyo número de neuronas, denotado como M, se selecciona mediante un proceso de CV realizado de manera independiente sobre cada problema de clasificación y sobre cada algo-

ritmo, ya sea el RA, el RA-we (con un valor de λ concreto) o el DW-RA; para el caso del algoritmo CV RA-we se seleccionan conjuntamente, mediante el proceso de CV, el valor de M y el de λ.

En cuanto a la función de coste para el entrenamiento, en el caso del algoritmo RA, los MLPs se entrenan con el objetivo de (2.3), i.e,

$$C_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2 \tag{6.1}$$

mientras que para los algoritmos RA-we, CV RA-we y DW-RA se minimiza la función (3.8),

$$C_{\lambda,t} = \sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}^{(l)})[d^{(l)} - o_t(\mathbf{x}^{(l)})]^2 \tag{6.2}$$

Para ello, se inicializan aleatoriamente las componentes de los pesos del MLP con valores uniformemente distribuidos en el intervalo $[-1, 1]$, y se realiza su actualización mediante el algoritmo de retropropagación. Los ajustes para este algoritmo son: un paso de aprendizaje inicial de 0.01 (tanto para la capa de entrada como para la de salida) que decrece linealmente hasta 0 durante las 100 épocas que dura el entrenamiento (habiéndose comprobado que éste es un número suficiente para asegurar la convergencia).

Además, se emplea un algoritmo de detención prematura del entrenamiento, utilizando el $80\,\%$ de los datos disponibles para entrenar el MLP y el $20\,\%$ restante para validar el diseño, es decir, para elegir los pesos correspondientes a la época que presenta un menor error sobre esta partición, reduciendo así los efectos perniciosos del sobreajuste.

Selección del número de clasificadores base

El último de los aspectos a considerar en el diseño es el número total de clasificadores empleados, es decir, el número total de rondas, T. Aunque este parámetro puede seleccionarse con técnicas de validación cruzada o atendiendo a la evolución del error de entrenamiento, ninguna de estas soluciones es adecuada para obtener una buena generalización. Así, por ejemplo, si se emplease un proceso de validación cruzada, se reduciría el número

efectivo de muestras de entrenamiento, modificando por ello la velocidad de convergencia del algoritmo y llegando a resultados de convergencia bastante distintos de los óptimos, lo que podría conducir a una inadecuada selección del punto de parada. Si, por el contrario, se atendiese al valor del error de entrenamiento, no habría garantías de que el error de generalización presente un comportamiento similar, ya que podría ocurrir que el error de entrenamiento dejase de reducirse o, incluso, llegase a anularse, y aún así, tal y como se explicó en el Apartado 2.2.2, podría ser conveniente aumentar el número de redes base del conjunto para mejorar su capacidad de generalización.

Por las razones anteriores, para el diseño de los conjuntos se ha optado por seleccionar el número final de rondas T en función de la evolución de los valores α_t, dado que cuando estos valores se hacen muy próximos a cero el añadir nuevos clasificadores al conjunto apenas afecta a las prestaciones del mismo. En otras palabras: se propone incrementar el número de redes hasta que la adición de nuevos componentes apenas tenga influencia en la función de clasificación. Concretamente, el criterio empleado consiste en detener el crecimiento del conjunto cuando el valor relativo de α_t en las últimas T_{ult} rondas es menor que un cierto valor de parada, C_{stop}; es decir,

$$\frac{\sum_{t'=T-T_{\text{ult}}+1}^{T} \alpha_{t'}}{T_{\text{ult}} \sum_{t'=1}^{T} \alpha_{t'}} < C_{\text{stop}} \tag{6.3}$$

fijando de forma experimental T_{ult} a 10 y C_{stop} a 0.01 para todos los algoritmos y todos los problemas, excepto, para *Ti*, que presenta una convergencia mucho más lenta, lo que aconseja utilizar $C_{\text{stop}} = 5 \cdot 10^{-6}$ para el algoritmo RA y $C_{\text{stop}} = 10^{-5}$ para el resto de algoritmos.

6.1.3. Diferencia estadística entre los resultados: T-test

Para evaluar las prestaciones de los diferentes algoritmos (RA, CV RA-we, comités de RA-we y DW-RA), se presentarán los valores medios, de 50 entrenamientos diferentes, de los errores de clasificación $\overline{E}_{clas}$ obtenidos por cada algoritmo junto con el tamaño medio de los conjuntos resultantes ($\overline{T}$); también se proporcionará la desviación estándar

de su estimación, es decir, la desviación estándar muestral divida por la raíz cuadrada del número de realizaciones.

Dado que para cada algoritmo y cada problema se tiene un conjunto de 50 valores distintos del error de clasificación, E_{clas}, para comparar los errores de clasificación entre dos algoritmos resulta interesante saber en qué casos sus poblaciones de muestras son estadísticamente diferentes; para ello, se aplicará el T-test [Hill y Lewicki, 2005]. Este test proporciona un valor, t, que resulta de calcular la diferencia entre las medias de las dos distribuciones que se comparan normalizada por sus desviaciones estándares; dado que se desconoce el tipo de distribución que siguen los grupos de datos a comparar, el T-test considera que sus medias siguen una distribución gaussiana de media $\{m_i\}_{i=1,2}$ (dada por la media muestral) y desviación estándar $\{\sigma_i/\sqrt{N_i}\}_{i=1,2}$, donde las desviaciones estándares $\{\sigma_i\}_{i=1,2}$ siguen una distribución Chi-cuadrado y los valores $\{N_i\}_{i=1,2}$ representan al número de realizaciones en cada grupo de datos. De este modo, el valor de t viene dado por

$$t = \frac{m_1 - m_2}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}} \tag{6.4}$$

y, tal y como se puede demostrar, se distribuye según una t de Student con $N_1 + N_2 - 1$ grados de libertad[1].

En el caso de los experimentos que se presentan en esta Tesis Doctoral, los valores m_1 y m_2 se corresponderán con los valores de $\overline{E}_{clas}$ de los algoritmos que se comparan y, dado que cada uno de estos valores se ha promediado sobre 50 iteraciones, el número de grados de libertad de distribución t de Student del parámetro t será 99. Entonces, y de acuerdo con las tablas de la distribución t de Student, se puede afirmar con un $95\,\%$ de seguridad que los valores medios de las dos distribuciones son estadísticamente diferentes si $|t| > 1.66$.

[1]El uso del T-test se está haciendo de manera laxa, ya para requiere que las realizaciones sean independientes; en los casos en los que se va a aplicar se emplean los mismos datos de entrenamiento, y aunque se se inicializan de forma distinta los pesos de los MLPs e, incluso, la partición de entrenamiento (consistente en el $80\,\%$ de los datos) es aleatoria, no se puede garantizar esta independencia.

Cuadro 6.2: Valores límite del parámetro t del T-test para distintos niveles de certeza.

Certeza ($\%$)	75	80	85	90	**95**	97.5	99	99.5	99.9	99.95
t_{limit}	0.677	0.845	1.042	1.290	**1.660**	1.984	2.364	2.626	3.174	3.390

A lo largo de la discusión de resultados, se utilizará $t_{\text{limit}} = 1.66$ como valor límite para aceptar diferencia estadística significativa entre los resultados alcanzados por dos algoritmos para cada problema. Pero, si se desea extraer conclusiones con un nivel de certeza distinto del $95\,\%$, podrían emplearse los valores límite correspondientes, algunos de los cuales se encuentran recogidos en el Cuadro 6.2.

6.2. SELECCIÓN POR CV DEL PARÁMETRO DE MEZCLA

En esta sección se comparan:

- El algoritmo RA básico descrito en el Capítulo 2; de cara a aclarar la notación, de ahora en adelante, este algoritmo se denotará como RA-se ("RA with standard emphasis"),
- y el algoritmo CV RA-we. Como se indicó en el Apartado 3.6, se trata de una versión particular del algoritmo RA-we en la que el valor del parámetro de mezcla, λ_{CV}, se selecciona mediante CV; concretamente, se ha realizado un proceso de CV de 5 particiones con 20 iteraciones sobre cada partición para seleccionar el parámetro de mezcla de entre 11 valores equiespaciados en el intervalo $[0, 1]$ (i.e., $\lambda_{CV} \in \{0, 0.1, 0.2, \ldots, 1\}$).

Además, se incluyen los resultados del algoritmo RA-we para el valor $\lambda_0 \in \{0, 0.1, 0.2, \ldots, 1\}$ que ofrece mejores resultados sobre el conjunto de test. Cierto es que este método para seleccionar λ no es lícito; sin embargo, esta aproximación "omnisciente" va a resultar de gran utilidad para evaluar las ventajas potenciales que el algoritmo RA-we

puede proporcionar y para comprobar la eficacia del método CV RA-we para la selección de λ.

En el Cuadro 6.3 se muestran los errores de clasificación, $\overline{E}_{clas}$, de ambos algoritmos junto con el tamaño medio ($\overline{T}$) de los conjuntos; también se indica, por un lado, el número de neuronas ocultas (M) que han empleado los MLPs que forman el conjunto (como se indicó, este valor ha sido seleccionado independientemente para cada algoritmo y problema mediante un proceso de CV), y, por otro lado, el valor proporcionado por el T-test aplicado a los errores de clasificación de los algoritmos RA-se y CV RA-we. Cabe reseñar que un valor positivo del T-test indica que la tasa de error del CV RA-we mejora la del RA-se, y un valor negativo indica lo contrario. Además, para facilitar la comparación entre los algoritmos, se ha resaltado con letra negrita aquel error de clasificación que resulta ser el menor de la comparación de estos dos algoritmos.

Del análisis de los resultados en el Cuadro 6.3 pueden extraerse las siguientes conclusiones:

- En dos de los problemas, *Ph* y *Ri*, el algoritmo CV RA-we mejora la tasa de error del RA-se con una diferencia estadística significativa ($t > 1.66$). En *Co*, el CV RA-we también obtiene una tasa de error menor, si bien la diferencia estadística no es tan clara. Analizando el resultado del T-test, y a la luz del Cuadro 6.2, se puede concluir que, con más de un $90\,\%$ de certeza, sí existe esta diferencia estadística.

- En otros tres casos (*Im*, *Kw* y *Sp*), ambos algoritmos presentan exactamente los mismos resultados dado que el algoritmo CV RA-we ha seleccionado 0.5 como valor del parámetro de mezcla.

- En los dos casos restantes, *Ab* y *Ti*, el CV RA-we obtiene peores tasas de error que el RA-se, aunque para el *Ti* no hay una diferencia estadística clara.

- Cabe resaltar que, aunque parezca que las mejoras proporcionadas por el algoritmo CV RA-we son pequeñas, realmente no lo son si se tiene en cuenta que las tasas de error que presenta el RA-se son de por sí bastante competitivas, ya que mejoran en

Cuadro 6.3: Prestaciones presentadas por los algoritmos RA-se y CV RA-we y por la aproximación "omnisciente".

	RA-se			CV RA-we				Aproximación "omnisciente"				T-test
	M	$\overline{T}$	$\overline{E}_{clas}$	M	λ_{CV}	$\overline{T}$	$\overline{E}_{clas}$	M	λ_0	$\overline{T}$	$\overline{E}_{clas}$	t
Ab	4	31.18 (± 0.36)	**19.38** (± 0.02)	4	0.8	33.06 (± 0.59)	19.45 (± 0.02)	6	0.1	18.70 (± 0.01)	19.01 (± 0.00)	-2.26
Co	2	33.68 (± 0.74)	29.00 (± 0.20)	6	0.1	26.04 (± 0.86)	**28.60** (± 0.21)	3	0	22.58 (± 0.07)	28.51 (± 0.02)	1.36
Im	11	19.60 (± 0.38)	**2.46** (± 0.04)	11	0.5	19.60 (± 0.38)	**2.46** (± 0.04)	9	0.3	19.72 (± 0.08)	2.26 (± 0.01)	0
Kw	15	29.26 (± 0.14)	**11.71** (± 0.01)	15	0.5	29.26 (± 0.14)	**11.71** (± 0.01)	9	0.4	25.56 (± 0.04)	11.64 (± 0.00)	0
Ph	60	27.74 (± 0.34)	14.04 (± 0.07)	62	0	16.56 (± 0.13)	**13.49** (± 0.09)	54	0.1	18.34 (± 0.02)	13.46 (± 0.01)	4.78
Ri	48	28.86 (± 0.18)	9.73 (± 0.01)	34	0.7	38.80 (± 1.05)	**9.64** (± 0.03)	34	1	47.68 (± 0.10)	9.30 (± 0.00)	3.05
Sp	7	26.18 (± 0.64)	**5.94** (± 0.09)	7	0.5	26.18 (± 0.64)	**5.94** (± 0.09)	6	0.6	33.04 (± 0.17)	5.77 (± 0.01)	0
Ti	4	4490.36 (± 139.66)	**0.78** (± 0.08)	3	0.4	6965.76 (± 161.41)	0.89 (± 0.08)	4	0.5	4490.36 (± 139.66)	0.78 (± 0.08)	-0.94

todos los casos, excepto en *Co*, las obtenidas por una SVM (véase el Cuadro 6.1); por lo tanto, la mejora que consigue el algoritmo CV RA-we es más relevante de lo que puede parecer a primera vista, más aún si se tiene en cuenta que la tasa de error proporcionada por las SVM en el problema *Co*, que no es batida por el RA-se, es alcanzada por el CV RA-we. En definitiva, se puede afirmar que el algoritmo CV RA-we mejora sistemáticamente las prestaciones de las SVM en todos las bases de datos empleadas en esta Tesis Doctoral.

Aparte de lo anterior, si se comparan los resultados del CV RA-we con la aproximación "omnisciente" para evaluar la idoneidad de seleccionar λ mediante CV, se puede concluir

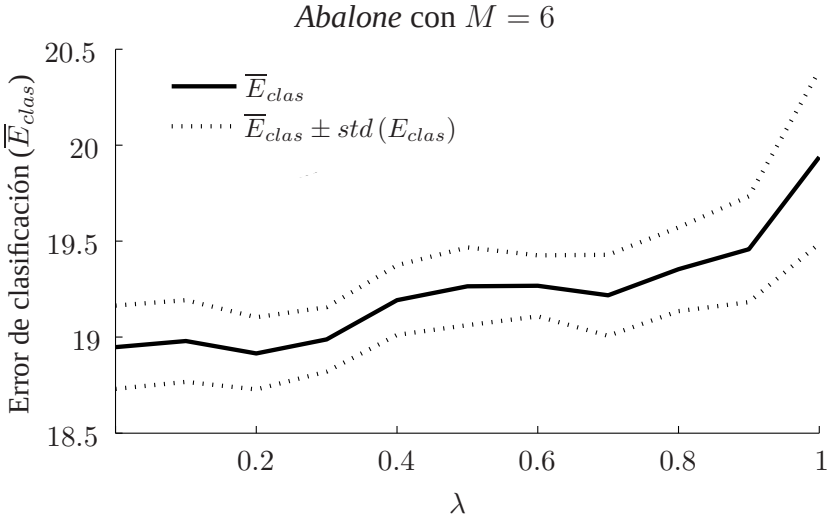

Figura 6.1: Comportamiento $\overline{E}_{clas}$ en función del valor de λ en el problema *Abalone*.

que el proceso de CV es únicamente moderadamente útil. Nótese que, aunque en la mayoría de los casos se ha seleccionado un valor próximo al óptimo, λ_0, una selección más fina en casos como *Im* , *Kw* o *Sp* proporcionaría mejoras todavía más significativas. Es más, hay casos como *Ab* en los que el valor λ_{CV} dista considerablemente de ser el óptimo; de hecho, en este caso se ha observado que existe un conjunto de valores próximos a $\lambda = 0.2$ que mejoran significativamente las prestaciones del CV RA-we (véase la Figura 6.1), no sólo en cuanto a tasa de error obtenida, sino también permitiendo la obtención de conjuntos de menor tamaño.

En definitiva, los resultados alcanzados muestran que, si bien el algoritmo CV RA-we presenta mejoras frente al RA-se en términos de reducción de la tasa de error y, en ocasiones, de reducción de la complejidad de la red resultante, no siempre es capaz de aprovechar al máximo las ventajas que la función de énfasis mixta es capaz de proporcionar.

6.3. COMITÉS DE CONJUNTOS RA-WE

A continuación se van a analizar las prestaciones los cuatro métodos de construcción de comités que se propusieron en el Capítulo 4, es decir:

1. Los comités construidos para minimizar el MSE y que emplean el método de combinación lineal. Por comodidad en la notación, de ahora en adelante este tipo de comités va a denotarse como comité lineal o $\mathrm{COM}_{\mathrm{lin}}$.

2. Los comités construidos para minimizar el MSE, pero que, a diferencia de los anteriores, emplean una función de activación de tipo tangente hiperbólica a la salida. En este caso, la notación empleada será comité con "tanh" o $\mathrm{COM}_{\mathrm{th}}$.

3. Los comités que utilizan el esquema de voto generalizado, a los que se citará como comité con voto o $\mathrm{COM}_{\mathrm{voto}}$.

4. Y, por último, los comités que siguen el criterio de ajuste de pesos del RA-we. Este tipo de comités se denotará con $\mathrm{COM}_{\mathrm{RA-we}}$.

Para construir cada uno de estos comités, se han entrenado 11 conjuntos tipo RA-we, empleando, cada uno de ellos, un valor diferente de λ de entre el conjunto de valores $\{0, 0.1, \ldots, 0.9, 1\}$, y seleccionando mediante un proceso de CV el número de neuronas ocultas (M) que deben emplear sus MLPs; esta selección, cuyos resultados se recogen en el Cuadro 6.4, se realiza para obtener el mejor conjunto RA-we y es, por lo tanto, independiente del tipo de comité que se vaya a construir.

Una vez que se dispone de los conjuntos que van a integrar el comité, se debe fijar el conjunto de pesos de salida $\{w_1, \ldots, w_{11}\}$ que realiza la combinación de los conjuntos RA-we. El comité lineal y el comité con voto emplean la inversa generalizada de Moore-Penrose para ajustar estos pesos, mientras que el comité con el criterio del RA-we sigue el procedimiento descrito en 4.1.1 y el comité con "tanh" requiere emplear un algoritmo de búsqueda para ajustar el conjunto de pesos. En particular, para entrenar la capa de salida de este último tipo de comités se emplea un algoritmo de descenso por gradiente que se

Cuadro 6.4: Valores de M seleccionados por cada conjunto RA-we en función del valor de λ empleado por cada uno.

	Valor de λ empleado										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Ab	4	4	5	3	5	4	6	2	4	5	3
Co	3	6	4	4	2	2	2	2	2	2	2
Im	15	14	14	15	11	11	11	11	10	14	10
Kw	15	13	13	14	15	15	13	12	12	13	13
Ph	62	64	64	60	60	60	66	66	58	66	58
Ri	30	30	44	36	36	48	48	34	34	34	34
Sp	5	5	5	6	7	7	5	7	8	8	5
Ti	3	4	3	3	3	4	4	3	3	4	3

deja evolucionar durante 100 épocas (suficiente para asegurar la convergencia), con una tasa de aprendizaje fijada a 0.05 durante las primeras 50 épocas y decreciente desde 0.05 a 0 durante las 50 épocas restantes.

Además de evaluar las prestaciones de cada tipo de comité, se analizarán las ventajas que aporta el método de selección de redes sobre cada uno de ellos. Dicho método mide la calidad de los diferentes conjuntos RA-we que van a formar el comité y elimina las redes RA-we que presentan las peores prestaciones y pueden deteriorar las prestaciones finales del comité; para ello, requiere establecer, previamente, un umbral que indique qué redes deben eliminarse, para cuya selección se ha empleado un proceso de CV (aplicado únicamente para la capa de salida) sobre el conjunto de parámetros de calidad normalizado (véase el Apartado 4.1.2). Los umbrales seleccionados para cada problema y cada tipo de comité se encuentran recogidos en el Cuadro 6.5.

En los apartados siguientes se presentarán las prestaciones de cada tipo de comité frente al algoritmo RA-se, comparando sus prestaciones cuando se emplea el método de selección de redes y cuando no es así (se llamará a esta versión como método básico). Aparte de los valores medios y las estimaciones de las desviaciones de las tasas de error

Cuadro 6.5: Umbrales empleados por el método de selección de redes en cada problema y en cada tipo de comité.

	$\mathrm{COM_{lin}}$	$\mathrm{COM_{th}}$	$\mathrm{COM_{voto}}$	$\mathrm{COM_{RA-we}}$
Ab	0.1	-0.8	0	1.2
Co	-1.1	-0.6	-1.4	-1.1
Im	-0.7	1.3	-0.6	-0.6
Kw	2.5	4.1	2.8	2.4
Ph	-0.6	-0.6	-0.5	-0.6
Ri	0.3	0.7	-0.4	0.5
Sp	-0.8	1.3	-0.7	-0.6
Ti	-0.3	-0.3	-0.3	-0.3

(E_{clas}) y del numero de redes (T), se incluyen los valores del T-test obtenidos al comparar los errores del comité básico con el RA-se ($t_{Bas,RA}$), al comparar el comité con selección de redes con el RA-se ($t_{Sel,RA}$) y al comparar entre sí el comité con selección de redes con el comité básico ($t_{Sel,Bas}$).

6.3.1. Prestaciones de los comités lineales

En el Cuadro 6.6 se presentan los resultados de este tipo de comités, de los cuales se concluye que:

- Los comités lineales básicos presentan unas prestaciones relativamente buenas, ya que en cinco de los ocho problemas sobre los que han sido evaluados (*Ab*, *Co*, *Kw*, *Ph* y *Sp*) se consigue mejorar las prestaciones del RA-se. Además, en cuatro de estos problemas se obtiene una diferencia estadística clara ($t_{Bas,RA} > 1.66$).

- Cuando se emplea el método de selección de redes sobre este tipo de comités, se mejoran sus prestaciones en la mayoría de los casos, destacando las mejoras en las tasas de error de *Ab*, *Im*, *Ri* y *Ti*, ya que presentan una diferencia estadística

Cuadro 6.6: Prestaciones de los comités lineales (en su versión básica y realizando selección de redes) frente al RA-se.

	RA-se		Básico			Con selección de redes			
	$\overline{T}$	$\overline{E}_{clas}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Bas,RA}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Sel,RA}$	$t_{Sel,Bas}$
Ab	31.18 (± 0.36)	19.38 (± 0.02)	310.48 (± 1.31)	19.31 (± 0.02)	2.41	160.86 (± 4.30)	**19.21** (± 0.02)	5.20	3.34
Co	33.68 (± 0.74)	29.00 (± 0.20)	321.58 (± 2.44)	28.68 (± 0.20)	1.12	48.52 (± 2.48)	**28.64** (± 0.18)	1.31	0.14
Im	19.60 (± 0.38)	2.46 (± 0.04)	226.48 (± 1.96)	2.65 (± 0.03)	−3.62	60.04 (± 2.70)	**2.35** (± 0.03)	2.11	7.31
Kw	29.26 (± 0.14)	11.71 (± 0.01)	336.34 (± 1.22)	**11.66** (± 0.00)	6.59	334.56 (± 1.77)	11.66 (± 0.00)	6.55	−0.18
Ph	27.74 (± 0.34)	14.04 (± 0.07)	292.22 (± 2.11)	13.73 (± 0.07)	3.17	62.88 (± 3.52)	**13.49** (± 0.09)	4.82	2.19
Ri	28.86 (± 0.18)	**9.73** (± 0.01)	348.62 (± 2.06)	9.97 (± 0.01)	−16.52	254.92 (± 4.46)	9.75 (± 0.02)	−0.88	11.28
Sp	26.18 (± 0.64)	5.94 (± 0.09)	309.84 (± 3.13)	**5.55** (± 0.07)	3.43	78.84 (± 2.97)	5.68 (± 0.08)	2.22	−1.16
Ti	4490.36 (± 139.66)	**0.78** (± 0.08)	48350.86 (± 757.91)	2.85 (± 0.19)	−10.20	26566.84 (± 633.52)	1.17 (± 0.12)	−2.80	7.64

muy significativa. Sin embargo, en los problemas *Sp* y *Kw*, este método degrada las prestaciones de los comités lineales, aunque, en el cualquier caso, la degradación sufrida es despreciable.

- De otro lado, cabe destacar la reducción de necesidades computacionales que el método de selección de redes aporta a los comités construidos, ya que llega a ser del $85\,\%$ en *Co*, del $80\,\%$ en *Ph* y del $75\,\%$ en *Im* y *Sp*.

- Por último, si se comparan los resultados del RA-se con los obtenidos por los comités empleando el método de selección de redes, se ve que estos comités mejoran las prestaciones del RA-se, además de en los mismos casos que la versión básica,

para el problema *Im*. En el problema *Co*, donde no hay diferencia estadística clara frente a los resultados del RA-se, el T-test está próximo a indicarla con una certeza del $90\,\%$. En los problemas *Ri* y *Ti*, los resultados siguen siendo peores que los del RA-se, pero ahora la diferencia es menor; de hecho, en *Ri* no hay una diferencia estadística clara entre ellos ($|t_{Sel,RA}| < 1.66$).

6.3.2. Prestaciones de los comités con activación "tanh"

Los resultados de este tipo de comité, recogidos en el Cuadro 6.7, muestran que:

- Este tipo de comité, en su versión básica, sólo mejora las prestaciones del RA-se, con diferencia estadística clara, en dos problemas: *Im* y *Ph*. En el resto de los casos las prestaciones son peores, salvo en *Ab*, donde puede considerarse que presenta los mismos resultados del RA-se.

- El empleo del método de selección de redes aporta ventajas en cuatro casos, aunque estas ventajas son mínimas: sólo en uno de estos casos, en el problema *Ti*, hay diferencia estadística con los resultados de la versión básica ($t_{Sel,Bas} = 6.58$). En *Ab*, la selección de redes permite que este tipo de comité mejore la tasa de error del RA-se (aunque sin diferencia estadística clara). En el resto de los problemas, el método de selección de redes empeora las prestaciones del comité. El motivo de esta degradación en la tasa de error puede deberse a una mala selección de los umbrales empleados para la selección de redes, ya que, como puede comprobarse en el Cuadro 6.5, los umbrales seleccionados para este comité difieren claramente de los umbrales seleccionados para el resto de comités.

- Obviamente, a pesar de esta degradación ocasional en la tasa de error, la aplicación del método de selección de redes sí consigue una reducción del número de redes que componen el comité, aunque no tan clara como en el caso anterior: en problemas como *Ab*, *Co*, *Ph* y *Ti* se consiguen reducciones entre el $50\,\%$ y el $80\,\%$.

Cuadro 6.7: Prestaciones de los comités con activación "tanh" (en su versión básica y realizando selección de redes) frente al RA-se.

	RA-se		Básico			Con selección de redes			
	$\overline{T}$	$\overline{E}_{clas}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Bas,RA}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Sel,RA}$	$t_{Sel,Bas}$
Ab	31.18 (± 0.36)	19.38 (± 0.02)	310.48 (± 1.31)	19.38 (± 0.05)	-0.02	63.08 (± 3.17)	**19.35** (± 0.04)	0.64	0.48
Co	33.68 (± 0.74)	**29.00** (± 0.20)	321.58 (± 2.44)	29.30 (± 0.22)	-0.98	85.86 (± 2.27)	29.18 (± 0.25)	-0.57	0.35
Im	19.60 (± 0.38)	2.46 (± 0.04)	226.48 (± 1.96)	**2.24** (± 0.03)	3.85	196.62 (± 2.06)	2.43 (± 0.06)	0.37	-2.92
Kw	29.26 (± 0.14)	**11.71** (± 0.01)	336.34 (± 1.22)	11.80 (± 0.01)	-7.25	336.34 (± 1.22)	11.78 (± 0.01)	-5.26	1.28
Ph	27.74 (± 0.34)	14.04 (± 0.07)	292.22 (± 2.11)	**13.38** (± 0.09)	5.77	62.88 (± 3.52)	13.51 (± 0.09)	4.73	-1.00
Ri	28.86 (± 0.18)	**9.73** (± 0.01)	348.62 (± 2.06)	9.96 (± 0.04)	-5.56	275.94 (± 4.58)	10.78 (± 0.16)	-6.45	-4.89
Sp	26.18 (± 0.64)	**5.94** (± 0.09)	309.84 (± 3.13)	6.22 (± 0.09)	-2.28	273.06 (± 3.46)	6.61 (± 0.10)	-4.97	-2.89
Ti	4490.36 (± 139.66)	**0.78** (± 0.08)	48350.86 (± 757.91)	3.23 (± 0.28)	-8.38	26566.84 (± 633.52)	1.23 (± 0.11)	-3.23	6.58

- Para finalizar, cabe destacar que aunque las prestaciones conseguidas por este tipo de comité no mejoren casi nunca a las del RA-se, en los casos en que sí lo hacen, se consigue una ganancia muy significativa que, como se verá más adelante, mejora incluso las prestaciones de la aproximación "omnisciente".

6.3.3. Prestaciones de los comités con voto generalizado

De los resultados presentados por los comités construidos con el método de voto generalizado, que se recogen en el Cuadro 6.8, se observa:

Cuadro 6.8: Prestaciones de los comités con voto generalizado (en su versión básica y realizando selección de redes) frente al RA-se.

	RA-se		Básico			Con selección de redes			
	$\overline{T}$	$\overline{E}_{clas}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Bas,RA}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Sel,RA}$	$t_{Sel,Bas}$
Ab	31.18 (± 0.36)	19.38 (± 0.02)	310.48 (± 1.31)	19.28 (± 0.02)	3.57	153.22 (± 4.38)	**19.19** (± 0.02)	6.33	3.45
Co	33.68 (± 0.74)	29.00 (± 0.20)	321.58 (± 2.44)	28.69 (± 0.21)	1.05	30.80 (± 1.96)	**25.68** (± 0.44)	6.81	6.17
Im	19.60 (± 0.38)	2.46 (± 0.04)	226.48 (± 1.96)	2.67 (± 0.03)	−3.90	76.32 (± 3.43)	**2.34** (± 0.03)	2.33	8.07
Kw	29.26 (± 0.14)	11.71 (± 0.01)	336.34 (± 1.22)	**11.66** (± 0.00)	5.90	336.34 (± 1.22)	**11.66** (± 0.00)	5.90	0.00
Ph	27.74 (± 0.34)	14.04 (± 0.07)	292.22 (± 2.11)	13.73 (± 0.06)	3.16	79.80 (± 3.45)	**13.50** (± 0.09)	4.70	2.12
Ri	28.86 (± 0.18)	9.73 (± 0.01)	348.62 (± 2.06)	9.96 (± 0.01)	−15.96	164.10 (± 6.60)	**9.52** (± 0.02)	7.89	18.34
Sp	26.18 (± 0.64)	5.94 (± 0.09)	309.84 (± 3.13)	**5.55** (± 0.07)	3.44	90.68 (± 2.80)	5.65 (± 0.08)	2.46	−0.92
Ti	4490.36 (± 139.66)	**0.78** (± 0.08)	48350.86 (± 757.91)	2.94 (± 0.19)	−10.54	26566.84 (± 633.52)	1.32 (± 0.12)	−3.61	7.19

- Este tipo de comité, en su versión básica, mejora las prestaciones del RA-se en cinco de los ocho problemas (*Ab*, *Co*, *Kw*, *Ph* y *Sp*) y, además, lo hace presentando una diferencia estadística significativa en todos los casos excepto en uno (*Co*).

- Cuando estos comités emplean el método de selección de redes, consiguen reducir aún más las tasas de error del RA-se. Únicamente en *Sp* hay un ligero empeoramiento respecto a su versión básica ($t_{Sel,Bas} = -0.92$). Entre estas mejoras destaca la correspondiente al problema *Co*, donde la reducción de la tasa de error es bastante considerable (compárese incluso con el resultado obtenido por una SVM, $28.61\,\%$).

  Si se analiza en detalle la tasa de error obtenida en *Co* empleando el método de

selección de redes, se observa que en algunas ocasiones (en algunas de las 50 realizaciones sobre las que se encuentran promediados los resultados) el comité resultante es realmente bueno, presentando tasas de error del $22\,\%$, mientras que en otras ocasiones el comité sigue prestando tasas de error en torno al $29\,\%$; esto origina que el valor medio del error disminuya, respecto a la versión básica del comité, y aumente su desviación estándar. Por este motivo, aunque la diferencia entre los errores medios sea superior al $2\,\%$ en términos absolutos, el T-test marca la una diferencia estadística menor que en *Ri* o *Im*.

- Respecto al tamaño de los comités, se observa que la selección de redes ofrece reducciones computacionales generalmente en torno al $60\,\%$: en problemas como *Kw* no hay reducción alguna, pero en otros casos, como *Co*, la reducción computacional obtenida llega al $90\,\%$.

- Las mejoras aportadas por el método de selección de redes hacen que este tipo de comité supere sistemáticamente al RA-se, con una diferencia estadística clara, salvo en el caso del problema *Ti*.

6.3.4. Prestaciones de los comités con criterio RA-we

Sobre este último tipo de comité, cuyos resultados se presentan en el Cuadro 6.9, se puede concluir que:

- En su versión básica consiguen mejoras respecto al RA-se en cinco de los ocho problemas considerados: concretamente, en *Ab*, *Co*, *Im*, *Ph* y *Sp*. Sin embargo, únicamente en dos de ellos (*Im* y *Ph*) existe una diferencia estadística significativa.

- El método de selección apenas ocasiona diferencias con respecto al método básico, pues sólo aparece una diferencia estadística significativa en el problema *Ri*. El hecho de que sólo existan pequeñas mejoras se debe a que este tipo de comités emplea el criterio de ajuste de pesos del RA-we, el cual tiene en cuenta, evaluando su

Cuadro 6.9: Prestaciones de los comités con el criterio del RA-we (en su versión básica y realizando selección de redes) frente al RA-se.

	RA-se		Básico			Con selección de redes			
	$\overline{T}$	$\overline{E}_{clas}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Bas,RA}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Sel,RA}$	$t_{Sel,Bas}$
Ab	31.18 (± 0.36)	19.38 (± 0.02)	310.48 (± 1.31)	**19.33** (± 0.02)	1.57	268.96 (± 2.66)	19.34 (± 0.02)	1.48	−0.09
Co	33.68 (± 0.74)	29.00 (± 0.20)	321.58 (± 2.44)	28.61 (± 0.19)	1.42	48.52 (± 2.48)	**28.60** (± 0.19)	1.44	0.03
Im	19.60 (± 0.38)	2.46 (± 0.04)	226.48 (± 1.96)	2.35 (± 0.03)	2.03	76.32 (± 3.43)	**2.33** (± 0.03)	2.56	0.72
Kw	29.26 (± 0.14)	**11.71** (± 0.01)	336.34 (± 1.22)	11.74 (± 0.01)	−1.90	333.64 (± 2.06)	11.74 (± 0.01)	−1.97	−0.01
Ph	27.74 (± 0.34)	14.04 (± 0.07)	292.22 (± 2.11)	13.49 (± 0.09)	4.68	62.88 (± 3.52)	**13.47** (± 0.09)	4.84	0.14
Ri	28.86 (± 0.18)	**9.73** (± 0.01)	348.62 (± 2.06)	9.87 (± 0.02)	−6.66	262.10 (± 4.83)	9.83 (± 0.02)	−4.47	1.68
Sp	26.18 (± 0.64)	5.94 (± 0.09)	309.84 (± 3.13)	5.80 (± 0.08)	1.21	108.18 (± 2.93)	**5.74** (± 0.08)	1.65	0.47
Ti	4490.36 (± 139.66)	**0.78** (± 0.08)	48350.86 (± 757.91)	0.97 (± 0.09)	−1.55	26566.84 (± 633.52)	0.91 (± 0.09)	−1.10	0.40

parámetro generalizado de separación, la relevancia de cada conjunto, asignando un peso muy bajo a aquellos con peores prestaciones.

- Sin embargo, el método de selección de redes sí consigue una reducción computacional significativa, como puede verse en *Co*, *Im*, *Ph*, *Sp* o *Ti*.

- Debido a que el método de selección de redes no aporta mejoras notables, la comparación entre el comité con selección de redes y el RA-se puede realizarse en términos muy similares a los del comité básico.

6.3.5. Evaluación conjunta de los diferentes comités

Hasta el momento se han analizado las prestaciones de cada tipo de comité frente al RA-se. En este apartado se va a llevar a cabo un evaluación conjunta de todos ellos. Por este motivo, en el Cuadro 6.10 se muestran los resultados del mejor de los comités resultantes de los apartados anteriores frente al RA-se; en el caso de que varios tipos de comités presenten los mismos errores de clasificación, se muestra el resultado de la combinación que tenga una mayor diferencia estadística frente al RA-se. Además, en el cuadro también se incluyen los resultados de la aproximación "omnisciente", para poder así evaluar en qué casos la combinación es capaz de superar al mejor de los conjuntos RA-we que constituyen el comité.

Los resultados del Cuadro 6.10 evidencian las ventajas que los comités de redes RA-we pueden aportar, ya que son capaces de mejorar las prestaciones del RA-se, con una diferencia estadística clara, en todos los casos excepto en *Ti*. Además, en la mitad de los casos, concretamente en los problemas *Co*, *Im*, *Ph* y *Sp*, se consigue reducir la tasa de error presentada por la aproximación "omnisciente", lo que indica la idoneidad del proceso de combinación, ya que es capaz de mejorar las prestaciones de la mejor de las redes RA-we que componen el comité. En *Co* y *Sp* esta tasa "record" corresponde a los comités con voto, mientras que en *Im* y *Ph* son los comités con "tanh" los que consiguen esta reducción de la tasa de error.

Adicionalmente, comparando entre sí los resultados incluidos en los Cuadros 6.6 a 6.10, se concluye que:

- Si se consideran las prestaciones medias (sobre todas las bases de datos), se puede afirmar que los comités lineales y los comités con voto son los que presentan los mejores resultados, destacando especialmente los comités con voto, ya que consiguen tasas de error record en cuatro de los ocho problemas.

- Es notable la mejora casi sistemática que el método de selección de redes ocasiona sobre los comités lineales y con voto, aportando mejoras en la tasa de error y reduciendo la complejidad del comité.

Cuadro 6.10: Evaluación del mejor de los comités frente al algoritmo RA-se y frente a la aproximación "omnisciente".

	RA-se		Comités			T-test	Omnisciente	
	$\overline{T}$	$\overline{E}_{clas}$	Tipo	$\overline{T}$	$\overline{E}_{clas}$	t	$\overline{T}$	$\overline{E}_{clas}$
Ab	31.18 (± 0.36)	19.38 (± 0.02)	$\mathrm{COM_{voto}}$ Selección	153.22 (± 4.38)	**19.19** (± 0.02)	6.33	18.70 (± 0.01)	19.01 (± 0.00)
Co	33.68 (± 0.74)	29.00 (± 0.20)	$\mathrm{COM_{voto}}$ Selección	30.80 (± 1.96)	**25.68** (± 0.44)	6.81	22.58 (± 0.07)	28.51 (± 0.02)
Im	19.60 (± 0.38)	2.46 (± 0.04)	$\mathrm{COM_{th}}$ Básico	226.48 (± 1.96)	**2.24** (± 0.03)	3.85	19.72 (± 0.08)	2.26 (± 0.01)
Kw	29.26 (± 0.14)	11.71 (± 0.01)	$\mathrm{COM_{lin}}$ Bás. ó Sel.	336.34 (± 1.22)	**11.66** (± 0.00)	6.59	25.56 (± 0.04)	11.64 (± 0.00)
Ph	27.74 (± 0.34)	14.04 (± 0.07)	$\mathrm{COM_{th}}$ Básico	292.22 (± 2.11)	**13.38** (± 0.09)	5.77	18.34 (± 0.02)	13.46 (± 0.01)
Ri	28.86 (± 0.18)	9.73 (± 0.01)	$\mathrm{COM_{voto}}$ Selección	164.10 (± 6.60)	**9.52** (± 0.02)	7.89	47.68 (± 0.10)	9.30 (± 0.00)
Sp	26.18 (± 0.64)	5.94 (± 0.09)	$\mathrm{COM_{voto}}$ Básico	309.84 (± 3.13)	**5.55** (± 0.07)	3.44	33.04 (± 0.17)	5.77 (± 0.01)
Ti	4490.36 (± 139.66)	**0.78** (± 0.08)	$\mathrm{COM_{RA-we}}$ Selección	26566.84 (± 633.52)	0.91 (± 0.09)	−1.10	4490.36 (± 139.66)	0.78 (± 0.08)

- Por otro lado, los comités con "tanh", en su versión básica, son capaces de presentar tasas de error inferiores a las de la aproximación "omnisciente" en los problemas *Im* y *Ph*, aunque en el resto de los problemas sus prestaciones dejan mucho que desear. Esta diferencia de calidad tan considerable se debe al proceso de ajuste de pesos que, en ocasiones, causa sobreajuste en el comité, deteriorando así su capacidad de generalización.

- El último tipo de comités, los comités con el criterio del RA-we, presenta tasas de error que superan en muchas ocasiones a las del RA-se, y que no distan considerablemente de los resultados del mejor comité; es más, en algún caso presenta la mejor

tasa de error de entre todos los comités.

En resumen, se concluye que, aunque los comités con "tanh" presenten resultados puntuales muy buenos, en general son mejores los resultados del resto de comités, destacando los comités lineales y, especialmente, los comités con voto cuando emplean el método de selección de redes.

6.4. CLASIFICACIÓN ACELERADA DE COMITÉS DE CONJUNTOS RA-WE

En esta sección se van a evaluar las ventajas aportadas por el método de clasificación acelerada, presentado en la Sección 4.2, sobre los comités lineales, los comités con "tanh" y los comités con el criterio del RA-we (no se considerarán comités con voto ya que, tal y como se explicó en el Apartado 4.2.1, su uso es incompatible con el método de clasificación rápida). En la sección anterior se evaluaron dos versiones diferentes de cada unos de estos comités, la versión básica y la versión con selección de redes. Dado que esta última presenta (generalmente) mejores prestaciones, se aplicará el método de clasificación acelerada únicamente sobre dicha versión. Debemos indicar, no obstante, que si se aplicase sobre sobre la versión básica (que implica un mayor número de redes) las mejoras conseguidas, en términos de reducción del coste computacional, superarían a las que aparecen a continuación.

Para analizar las mejoras aportadas por el método de clasificación acelerada se presentan, para cuatro valores diferentes del parámetro suavizador del umbral (β), el valor medio del número de clasificadores base que es necesario evaluar ($\overline{T}^{\beta}$), promediado sobre todo el conjunto de test, y el error medio de clasificación conseguido ($\overline{E}^{\beta}_{clas}$). Además, junto con los valores medios se indicarán (entre paréntesis) los valores correspondientes a la reducción (media) relativa del tamaño del comité, dada por:

$$\Delta\overline{T}(\,\%) = \frac{\overline{T} - \overline{T}^{\beta}}{\overline{T}}$$

y a la reducción (media) relativa del error de clasificación:

$$\Delta\overline{E}_{clas}(\,\%) = \frac{\overline{E}_{clas} - \overline{E}^{\beta}_{clas}}{\overline{E}_{clas}}$$

donde $\overline{T}$ y $\overline{E}_{clas}$ representan los valores medios originales del número de máquinas y del error de clasificación, respectivamente. Nótese que cuando el signo de $\Delta\overline{E}_{clas}(\,\%)$ es negativo indica que se trata de un aumento del error.

Para facilitar la lectura de los resultados, para cada problema, se ha resaltado con letra negrita el caso (correspondiente al valor de β que se indica) que permite obtener la máxima reducción computacional mientras que se consigue la menor tasa de error posible.

6.4.1. Clasificación acelerada con comités lineales

El Cuadro 6.11 muestra las consecuencias de aplicar el método de clasificación rápida sobre los comités lineales. Se evidencia que el método de clasificación acelerada puede aportar considerables reducciones del coste computacional, pero que el valor de β más adecuado depende del problema que se considere. Así, por ejemplo:

- En problemas como *Im*, *Ph* y *Sp* deben usarse valores conservadores de β (sobre 0.5), consiguiendo con ellos reducciones computaciones en torno al $90\,\%$. En la Figura 6.2 se muestra, para los problemas *Im* y *Sp*, el comportamiento del error de clasificación y del número de máquinas evaluadas en función de β. Puede observarse que, para β decreciente de 1 a 0.5, el error de clasificación se mantiene constante, mientras que el número de máquinas se reduce. Si se empleasen valores todavía menores de β, el error de clasificación comenzaría a aumentar, degradándose las prestaciones del comité.

- En casos como *Ab*, *Co* y *Ri* se pueden usar valores menores de β ($\beta = 0.3$) sin deteriorar por ello la tasa de error y consiguiendo reducciones computacionales entre el $84\,\%$ y el $91\,\%$.

Cuadro 6.11: Prestaciones del método de clasificación acelerada en los comités lineales.

	$\beta = 1.0$		$\beta = 0.5$		$\beta = 0.3$		$\beta = 0.1$	
	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)
Ab	49.21 (69.41 %)	19.21 (0.00 %)	30.41 (81.10 %)	19.21 (0.00 %)	**20.73** (87.11 %)	**19.21** (0.01 %)	8.91 (94.46 %)	19.24 (-0.16 %)
Co	16.47 (66.06 %)	28.64 (0.00 %)	11.08 (77.16 %)	28.64 (-0.01 %)	**8.00** (83.51 %)	**28.64** (0.01 %)	3.63 (92.51 %)	28.66 (-0.08 %)
Im	9.09 (84.85 %)	2.35 (0.00 %)	**5.64** (90.60 %)	**2.35** (-0.18 %)	3.64 (93.93 %)	2.40 (-2.40 %)	1.71 (97.15 %)	2.94 (-25.46 %)
Kw	83.34 (75.09 %)	11.66 (0.00 %)	41.79 (87.51 %)	11.66 (0.00 %)	26.13 (92.19 %)	11.66 (-0.00 %)	**10.37** (96.90 %)	**11.64** (0.14 %)
Ph	11.61 (81.54 %)	13.49 (0.00 %)	**6.75** (89.26 %)	**13.49** (0.01 %)	4.74 (92.46 %)	13.50 (-0.09 %)	2.04 (96.76 %)	13.61 (-0.86 %)
Ri	74.06 (70.95 %)	9.75 (0.00 %)	40.16 (84.24 %)	9.75 (0.00 %)	**22.95** (91.00 %)	**9.75** (0.00 %)	7.81 (96.94 %)	9.84 (-0.98 %)
Sp	13.89 (82.38 %)	5.68 (0.00 %)	**8.22** (89.57 %)	**5.68** (-0.06 %)	5.30 (93.28 %)	5.72 (-0.73 %)	2.02 (97.43 %)	6.13 (-8.02 %)
Ti	23839.06 (10.27 %)	1.17 (0.00 %)	17696.68 (33.39 %)	1.21 (-2.67 %)	**12613.31** (52.52 %)	**1.10** (6.22 %)	2053.35 (92.27 %)	3.65 (-210.67 %)

- En los problemas *Kw* y *Ti* se logran ligeras ventajas en términos de reducción de la tasa de error para $\beta = 0.1$ y $\beta = 0.3$, respectivamente. Si se representa, para el problema *Kw*, la evolución del error de clasificación medio en función del número de clasificadores que son evaluados, se puede observar (véase la Figura 6.3) que los últimos clasificadores están causando un ligero sobreajuste. Por este motivo, y dado que cuando el método de clasificación acelerada emplea valores pequeños de β consigue evitar que sean evaluados los últimos clasificadores, se palia el efecto del sobreajuste y disminuye el error de clasificación.

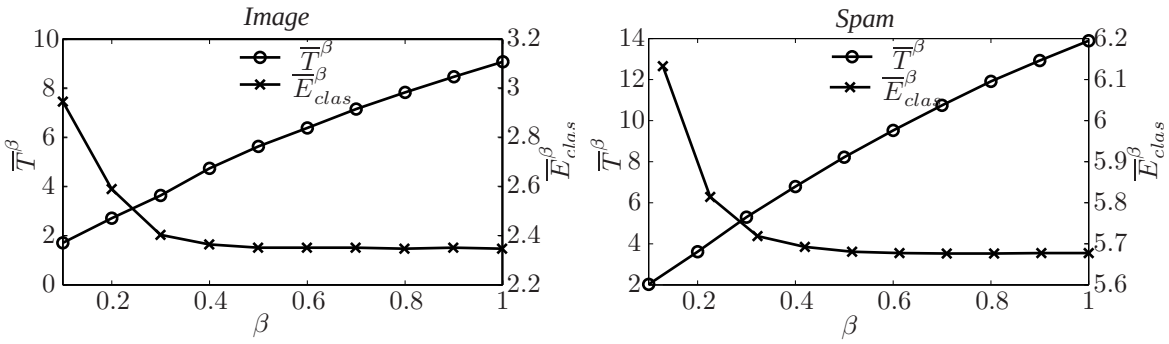

Figura 6.2: Evolución de $\overline{E}_{clas}$ y $\overline{T}$ en función de β en los comités lineales para los problemas *Image* y *Spam*.

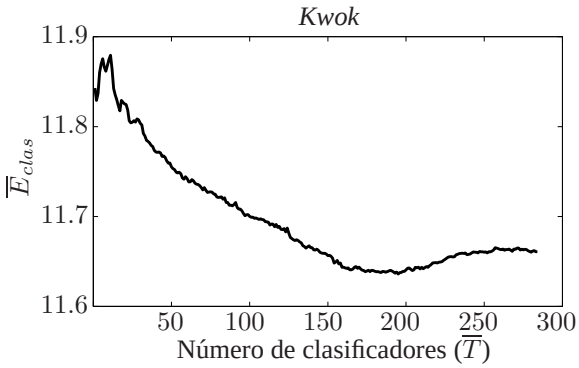

Figura 6.3: Evolución de $\overline{E}_{clas}$ en función del número de clasificadores evaluados por el método de clasificación acelerada en los comités lineales y el problema *Kwok*.

En definitiva, aunque la selección de β marca claramente las ventajas que se pueden obtener, para los comités lineales el empleo de un valor de β alrededor de 0.5 ofrecería, para todos los problemas, mejoras considerables en términos de reducción del coste computacional sin empeorar significativamente la tasa de error.

6.4.2. Clasificación acelerada con los comités con "tanh"

Tal y como muestran los resultados recogidos en el Cuadro 6.12, en este tipo de comités se pueden usar valores más agresivos de β que los sugeridos para los comités lineales,

Cuadro 6.12: Prestaciones del método de clasificación acelerada en los comités con "tanh".

	$\beta = 1.0$		$\beta = 0.5$		$\beta = 0.3$		$\beta = 0.1$	
	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)
Ab	36.04 (42.87 %)	19.35 (0.00 %)	22.15 (64.88 %)	19.35 (-0.01 %)	14.53 (76.96 %)	19.35 (-0.04 %)	**5.57** (91.17 %)	**19.32** (0.14 %)
Co	60.63 (29.38 %)	29.18 (0.00 %)	35.37 (58.81 %)	29.10 (0.28 %)	22.88 (73.35 %)	28.98 (0.69 %)	**8.52** (90.08 %)	**28.73** (1.57 %)
Im	45.57 (76.82 %)	2.43 (0.00 %)	20.48 (89.58 %)	2.25 (7.47 %)	**12.04** (93.88 %)	**2.22** (8.72 %)	3.45 (98.25 %)	2.54 (-4.45 %)
Kw	180.88 (46.22 %)	11.78 (0.00 %)	85.72 (74.51 %)	11.78 (0.00 %)	**48.68** (85.53 %)	**11.78** (0.00 %)	16.23 (95.17 %)	11.79 (-0.03 %)
Ph	14.45 (77.01 %)	13.51 (0.00 %)	**8.30** (86.80 %)	**13.50** (0.05 %)	5.34 (91.51 %)	13.52 (-0.08 %)	2.04 (96.75 %)	13.58 (-0.54 %)
Ri	207.01 (24.98 %)	10.78 (0.00 %)	109.38 (60.36 %)	9.45 (12.34 %)	62.35 (77.40 %)	9.47 (12.14 %)	**17.38** (93.70 %)	**9.28** (13.85 %)
Sp	86.48 (68.33 %)	6.61 (0.00 %)	36.80 (86.52 %)	6.03 (8.77 %)	20.55 (92.47 %)	5.79 (12.31 %)	**5.06** (98.15 %)	**5.76** (12.74 %)
Ti	**24139.90** (9.14 %)	**1.23** (0.00 %)	17955.37 (32.41 %)	1.26 (-2.54 %)	13150.70 (50.50 %)	1.67 (-35.59 %)	3176.72 (88.04 %)	8.72 (-607.20 %)

consiguiendo, además, reducciones en las tasas de error bastante más importantes que las obtenidas anteriormente. Así, por ejemplo:

- En los problemas *Ab*, *Co*, *Ri* y *Sp*, un valor de $\beta = 0.1$ no sólo consigue reducciones en la complejidad del comité del $90\,\%$ sino que, además, el error de clasificación disminuye, alcanzándose reducciones de hasta el $13\,\%$ y el $14\,\%$ para *Ri* y *Sp*, respectivamente. De hecho, el error de clasificación conseguido en *Ri* es menor que los ofrecidos por el resto de comités e, incluso, que el presentado por la aproximación "omnisciente".

  De nuevo, estas mejoras están causadas por la robustez que el método de clasificación acelerada proporciona frente al sobreajuste. Esta robustez es más beneficiosa

en los comités con "tanh" ya que, como se vio en el Apartado 6.3.2, estas máquinas son más proclives al problema de sobreajuste.

- En *Im*, *Kw* y *Ph*, es preferible emplear valores más elevados de β ($\beta = 0.3$ ó $\beta = 0.5$), pero aún así las mejoras son considerables, destacando la presentada en *Im*, problema en el que el coste computacional se reduce un $94\,\%$ y la tasa de error pasa a ser del $2.22\,\%$ (menor que la de cualquier otro comité y que la de la aproximación "omnisciente").

- Distinto es el caso del problema *Ti*, en el que se debe ser lo más conservador posible y emplear un valor de $\beta = 1$ si se desea evitar un empeoramiento en la tasa de error. Utilizando este valor de β se consiguen reducciones computacionales del $10\,\%$.

En conclusión, para este tipo de comité resulta generalmente más conveniente emplear valores de β entre 0.1 y 0.3, ya que de ese modo se reduce el efecto de sobreajuste y se consiguen, junto con las reducciones computacionales, mejoras en la tasa de error que hacen de los comités con "tanh" una alternativa válida a los comités lineales y con voto.

6.4.3. Clasificación acelerada de los comités con criterio RA-we

En este tipo de comité el método de clasificación acelerada da lugar a dos comportamientos claramente diferenciados:

- Por un lado, en los problemas *Im*, *Ph* y *Sp*, el método de clasificación rápida precisa una aplicación conservadora, ya que requiere el empleo de valores elevados de β (entre 0.5 y 1) para evitar deterioros. Para estos valores de β se consiguen reducciones computacionales de entre el $85\,\%$ y el $90\,\%$.

- Por otro lado, para los demás problemas, una aplicación más agresiva del método resulta beneficiosa. Así, se pueden emplear valores pequeños de β, por ejemplo

Cuadro 6.13: Prestaciones del método de clasificación acelerada en los comités con criterio RA-we.

	$\beta = 1.0$		$\beta = 0.5$		$\beta = 0.3$		$\beta = 0.1$	
	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)	$\overline{T}^{\beta}$ ($\Delta\overline{T}$)	$\overline{E}^{\beta}_{clas}$ ($\Delta\overline{E}_{clas}$)
Ab	45.82 (82.97 %)	19.34 (0.00 %)	26.72 (90.07 %)	19.34 (0.00 %)	17.62 (93.45 %)	19.34 (0.00 %)	**7.15** (97.34 %)	**19.33** (0.01 %)
Co	16.16 (66.69 %)	28.60 (0.00 %)	10.80 (77.73 %)	28.60 (-0.01 %)	**7.77** (83.99 %)	**28.59** (0.02 %)	3.47 (92.84 %)	28.62 (-0.07 %)
Im	**8.25** (89.20 %)	**2.33** (0.00 %)	5.00 (93.45 %)	2.35 (-0.74 %)	3.36 (95.59 %)	2.38 (-2.04 %)	1.20 (98.43 %)	2.81 (-20.63 %)
Kw	26.47 (92.07 %)	11.74 (0.00 %)	13.61 (95.92 %)	11.74 (0.00 %)	**8.43** (97.47 %)	**11.74** (0.00 %)	3.11 (99.07 %)	11.74 (-0.06 %)
Ph	**9.12** (85.50 %)	**13.47** (0.00 %)	5.38 (91.45 %)	13.48 (-0.03 %)	3.65 (94.19 %)	13.49 (-0.10 %)	1.57 (97.51 %)	13.56 (-0.64 %)
Ri	41.19 (84.29 %)	9.83 (0.00 %)	20.89 (92.03 %)	9.83 (0.00 %)	**12.29** (95.31 %)	**9.83** (0.00 %)	3.69 (98.59 %)	9.83 (-0.06 %)
Sp	11.94 (88.97 %)	5.74 (0.00 %)	**6.94** (93.59 %)	**5.74** (0.09 %)	4.34 (95.99 %)	5.75 (-0.13 %)	1.31 (98.79 %)	6.19 (-7.78 %)
Ti	17853.61 (32.80 %)	0.91 (0.00 %)	7753.19 (70.82 %)	0.91 (0.00 %)	3623.48 (86.36 %)	0.73 (20.00 %)	**692.68** (97.39 %)	**0.44** (52.00 %)

$\beta = 0.3$, para *Co*, o incluso, $\beta = 0.1$ en *Ab*, *Kw*, *Ri* y *Ti*, y conseguir con ello reducciones computacionales del $98\,\%$, a la vez que se mantiene la tasa de error constante en todos los casos. Sobresaliente resulta el caso del problema *Ti*, en el que el error de clasificación llega a ser de 0.44, batiendo así la tasa de error "record" que tenía el RA-se (0.78). De nuevo, esta mejora en la tasa de error se debe a que el método de clasificación acelerada palía el efecto de sobreajuste que afecta a este tipo de comités en el problema *Ti* (véase la Figura 6.4).

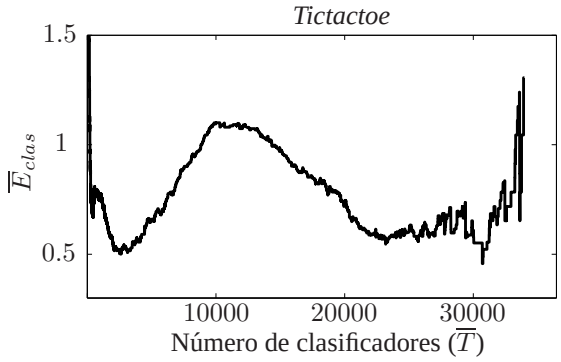

Figura 6.4: Evolución de $\overline{E}_{clas}$ en función del número de clasificadores evaluados por el método de clasificación acelerada en los comités criterio RA-we y el problema *Tictactoe*.

Aunque en este tipo de comités no resulta tan inmediata la selección del valor más adecuado de β, se puede elegir $\beta = 0.5$ sin deteriorar las prestaciones originales y consiguiendo reducciones computacionales importantes.

A la vista de los resultados conseguidos con los distintos tipos de comité, se puede concluir que el método de clasificación rápida disfruta de dos importantes ventajas: una esperada reducción del coste computacional durante la fase de test o fase operacional del sistema, junto con un aliviamiento del problema de sobreajuste presentado por algunos de los comités de conjuntos RA-we.

A la vista de los resultados alcanzados, queda claro que seleccionando un valor de $\beta = 1$ o, incluso, $\beta = 0.5$, se pueden conseguir ventajas sistemáticas en términos de ahorro computacional, afectando mínimamente a las prestaciones originales de la máquina en todos los problemas mencionados. Lógicamente, para aprovechar al máximo estas ventajas sería necesario seleccionar adecuadamente el valor de β, y, aunque ya se ha comprobado que para conjuntos RA-se puede emplearse un proceso de CV [Arenas-García et al., 2007], queda abierta esta línea de trabajo para, verdaderamente, poder explotar todo el potencial del método de clasificación acelerada sobre comités de conjuntos RA-we.

6.5. PRESTACIONES DEL ALGORITMO DW-RA

En esta sección se van a analizar las prestaciones del último de los algoritmos propuestos en la Tesis Doctoral, el DW-RA. Para ello, se compararán sus prestaciones con las de los algoritmos RA-se y CV RA-we, analizando no sólo las mejoras que el DW-RA puede aportar en términos de reducción del error de clasificación, sino también su velocidad de convergencia y su capacidad de generalización, de modo que se obtengan evidencias experimentales que corroboren los resultados teóricos presentados en la Sección 5.2. Para finalizar esta sección, se discutirán otras características de estos algoritmos, como su robustez frente a la selección del número de neuronas ocultas de los MLPs o la idoneidad del criterio de parada.

6.5.1. Una primera evaluación

En primer lugar, se van a analizar las ventajas que el DW-RA muestra respecto al RA-se y al CV RA-we, para lo que se ofrecen en el Cuadro 6.14 las estimaciones (promediadas sobre 50 iteraciones) de los errores de clasificación presentados por cada algoritmo ($\overline{E}_{clas}$) y el número de clasificadores base ($\overline{T}$) que componen cada conjunto, además de las desviaciones estándares de estas estimaciones (entre paréntesis). Por otro lado, se indica el número de neuronas ocultas (M) que emplean los MLPs que constituyen cada conjunto, y, por último, los indicadores de la diferencia estadística entre los errores de clasificación: los valores de t_{RA} y t_{CV} resultantes de aplicar el T-test a las tasas de error del DW-RA y el RA-se y del DW-RA y el CV RA-we, respectivamente.

Considerando las tasas de error ofrecidas por cada algoritmo, se observa cómo el DW-RA supera al RA-se en todos los problemas con una diferencia estadística clara, excepto en *Ti* (donde obtienen prestaciones muy similares). Si se compara con el CV RA-we, se observa cómo el DW-RA es ventajoso en los ocho problemas evaluados, aunque no siempre lo hace con una diferencia estadística clara (t_{CV} es menor que 1.66 en *Co*, *Ph* y *Ti*).

Cuadro 6.14: Prestaciones de los algoritmos RA-se, CV RA-we y DW-RA cuando cada algoritmo selecciona mediante un proceso CV independiente el valor de M empleado.

	RA-se			CV RA-we				DW-RA			T-test	
	M	$\overline{T}$	$\overline{E}_{clas}$	M	λ_{CV}	$\overline{T}$	$\overline{E}_{clas}$	M	$\overline{T}$	$\overline{E}_{clas}$	t_{RA}	t_{CV}
Ab	4	31.18 (± 0.36)	19.38 (± 0.02)	4	0.8	33.06 (± 0.59)	19.45 (± 0.02)	5	37.74 (± 0.28)	**18.97** (± 0.02)	14.45	15.42
Co	2	33.68 (± 0.74)	29.00 (± 0.20)	6	0.1	26.04 (± 0.86)	28.60 (± 0.21)	2	45.80 (± 0.62)	**28.54** (± 0.18)	1.71	0.24
Im	11	19.60 (± 0.38)	2.46 (± 0.04)	11	0.5	19.60 (± 0.38)	2.46 (± 0.04)	9	31.26 (± 0.37)	**2.31** (± 0.04)	2.51	2.51
Kw	15	29.26 (± 0.14)	11.71 (± 0.01)	15	0.5	29.26 (± 0.14)	11.71 (± 0.01)	22	31.62 (± 0.28)	**11.66** (± 0.01)	4.99	4.99
Ph	60	27.74 (± 0.34)	14.04 (± 0.07)	62	0.0	16.56 (± 0.13)	13.49 (± 0.09)	70	38.20 (± 0.69)	**13.43** (± 0.09)	5.27	0.48
Ri	48	28.86 (± 0.18)	9.73 (± 0.01)	34	0.7	38.80 (± 1.05)	9.64 (± 0.03)	44	36.02 (± 0.44)	**9.41** (± 0.03)	11.15	6.11
Sp	7	26.18 (± 0.64)	5.94 (± 0.09)	7	0.5	26.18 (± 0.64)	5.94 (± 0.09)	8	36.62 (± 0.55)	**5.75** (± 0.07)	1.66	1.66
Ti	4	4490.36 (± 139.66)	**0.78** (± 0.08)	3	0.4	6965.76 (± 161.41)	0.89 (± 0.08)	4	5698.42 (± 147.69)	0.79 (± 0.08)	-0.09	0.86

Por otro lado, si se examina el número medio de máquinas que requiere cada algoritmo, se observa que en la mayoría de los casos el algoritmo DW-RA requiere un número mayor de clasificadores base. Como se puede comprobar en la Figura 6.5, donde se muestra la evolución del error de clasificación para cada algoritmo, este mayor número de clasificadores base se debe al hecho de que el algoritmo DW-RA converge a tasas de error más baja, y no a una velocidad de convergencia más lenta.

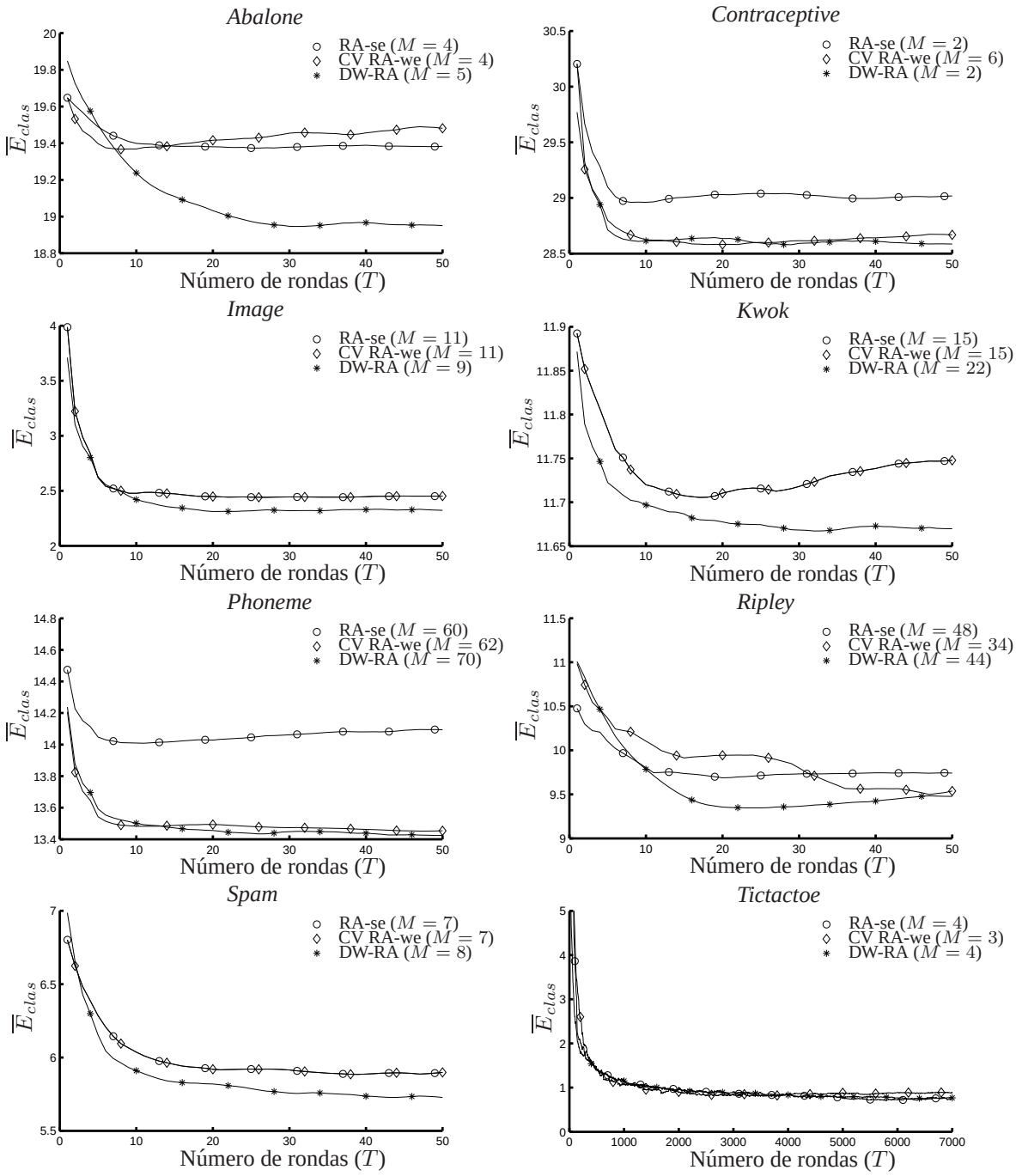

Figura 6.5: Convergencia de $\overline{E}_{clas}$ en los algoritmos RA-se, CV RA-we y DW-RA, en los ocho problemas considerados, cuando cada algoritmo selecciona independiente mediante CV el valor de M empleado.

6.5.2. Análisis de la velocidad de convergencia

Como era de esperar a partir del análisis teórico presentado en el Apartado 5.2.1, el algoritmo DW-RA suele presentar velocidades de convergencia más rápidas que las conseguidas por los otros dos algoritmos, tal y como se observa en la Figura 6.5 , de forma especialmente clara para los problemas *Ab*, *Kw*, *Ri* o *Sp* (en el resto de los problemas la convergencia es similar); además, nótese que en los problemas *Ab* y *Sp* el algoritmo DW-RA comienza con una tasa de error superior a la del RA-se y la del CV RA-we (debido a que usa un valor de M diferente) y, debido a su convergencia más rápida, en pocas iteraciones presenta tasas de error menores.

Para corroborar de manera más clara el resultado teórico al que se llegó en el Apartado 5.2.1, debe analizarse la evolución de la cota del error de entrenamiento (B_t), ya que en dicho apartado se sugirió que el algoritmo DW-RA es capaz de obtener una reducción más rápida de esta cota. En la Figura 6.6 se muestra la evolución del valor medio, promediado sobre 50 iteraciones, de la cota del error de entrenamiento ($\overline{B}_t$) para cada algoritmo cuando todos ellos emplean un mismo valor de M (de modo que la comparación no esté influida por la selección de este parámetro); concretamente, el valor de M empleado para esta comparación es el seleccionado mediante CV para el algoritmo RA-se. Como se observa, en casi todos los casos el DW-RA presenta, claramente, una convergencia más rápida, mostrando en todo momento valores menores de $\overline{B}_t$; sin embargo, hay tres casos en los que no es tan clara esta ventaja: en *Co*, donde el DW-RA presenta una velocidad de convergencia muy similar al CV RA-we; en *Ph*, donde el CV RA-we tiene una covergencia inicial mayor, aunque un valor final menor; y en *Ti*, donde su velocidad de convergencia es similar a la del RA-se y el CV RA-we. En resumen, se puede afirmar que la selección dinámica de λ que realiza el algoritmo DW-RA permite obtener velocidades de convergencia de $\overline{B}_t$ superiores y, en el peor de los casos, similares a las obtenidas por los algoritmos RA-se o CV RA-we.

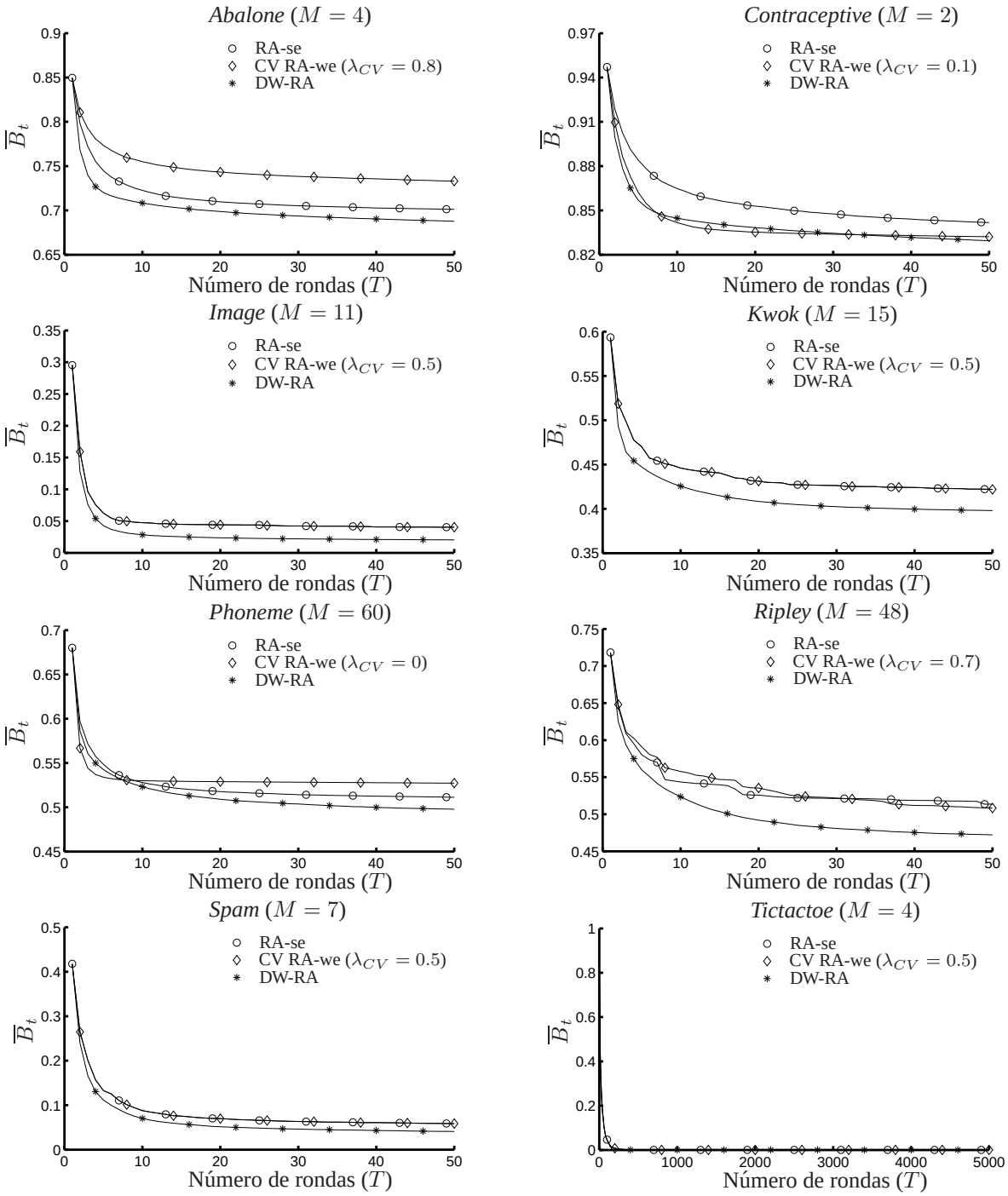

Figura 6.6: Convergencia de $\overline{B}_t$ en los algoritmos RA-se, CV RA-we y DW-RA, en los ocho problemas considerados, cuando el valor de M es el seleccionado por el RA-se.

6.5.3. Análisis de la capacidad de generalización

La segunda de las propiedades del algoritmo DW-RA que se analizó teóricamente fue su capacidad de generalización. En ese análisis se mostró que cabe esperar que el algoritmo DW-RA sea más eficiente que RA-we y RA-se a la hora de minimizar el riesgo marginal y, por lo tanto, presente un menor error de generalización. Para comprobar experimentalmente este resultado mediante una comparación equilibrada, se ha fijado el mismo M para los tres algoritmos, seleccionando, de nuevo, el valor de M obtenido por CV para el algoritmo RA-se. De esta manera, la complejidad de los MLPs es la misma para todos los algoritmos.

Como primera evidencia de la buena capacidad de generalización del algoritmo DW-RA, están los errores de clasificación que presentan los algoritmos cuando todos emplean el mismo valor de M. Tal y como se observa en el Cuadro 6.15, donde se recogen estos valores, los errores de clasificación obtenidos por el DW-RA son menores, en todos los casos excepto en *Ti*, a los correspondientes al RA-se y al CV RA-we.

Por otro lado, y para comprobar directamente que es esperable que el algoritmo DW-RA sea más eficiente en la minimización del riesgo marginal que los algoritmos RA-we y RA-se, también se ha analizado el comportamiento del valor medio del riesgo marginal, $\overline{R}_T^{\text{margin}}(\theta)$, en los tres algoritmos (fijando también el valor de M al seleccionado por el RA-se). Como resultado, se ha observado que $\overline{R}_T^{\text{margin}}(\theta)$ presenta, en la mayoría de los problemas, un comportamiento similar para los tres algoritmos, excepto en *Ri* donde se observa claramente (véase la Figura 6.7) que el valor del riesgo marginal es menor para el algoritmo DW-RA que para el resto de los algoritmos, con independencia del valor de θ considerado.

Para estudiar con mayor detalle el comportamiento del riesgo marginal, se han analizado sus valores en los alrededores de 0 y, concretamente, en el valor medio del margen mínimo ($\rho_{min} = \text{mín}_{l=1,\ldots,L}\, \rho_l$); así, por ejemplo, en el problema *Im* (véase la Figura 6.8(a) donde se muestra $\overline{R}_T^{\text{margin}}$ en las proximidades de 0) el valor medio de ρ_{min} es cercano a -0.2 en los algoritmos RA-se y CV RA-we, mientras que en el algoritmo DW-RA tiene un valor de -0.1. Es más, si se analiza la evolución de valor del valor medio de ρ_{min} du-

Cuadro 6.15: Prestaciones de los algoritmos RA-se, CV RA-we y DW-RA cuando el valor de M se ha fijado al seleccionado por el RA-se.

	M	RA-se		CV RA-we			DW-RA		T-test	
		$\overline{T}$	$\overline{E}_{clas}$	λ_{CV}	$\overline{T}$	$\overline{E}_{clas}$	$\overline{T}$	$\overline{E}_{clas}$	t_{RA}	t_{CV}
Ab	4	31.18 (± 0.36)	19.38 (± 0.02)	0.8	33.06 (± 0.59)	19.45 (± 0.02)	37.45 (± 0.32)	**19.09** (± 0.02)	9.45	10.86
Co	2	33.68 (± 0.74)	29.00 (± 0.20)	0.1	24.94 (± 0.39)	28.85 (± 0.21)	45.80 (± 0.62)	**28.54** (± 0.18)	1.71	1.13
Im	11	19.60 (± 0.38)	2.46 (± 0.04)	0.5	19.60 (± 0.38)	2.46 (± 0.04)	29.74 (± 0.48)	**2.35** (± 0.04)	1.83	1.83
Kw	15	29.26 (± 0.14)	11.71 (± 0.01)	0.5	29.26 (± 0.14)	11.71 (± 0.01)	32.46 (± 0.23)	**11.70** (± 0.01)	1.91	1.91
Ph	60	27.74 (± 0.34)	14.04 (± 0.07)	0.0	16.98 (± 0.14)	13.73 (± 0.08)	37.63 (± 0.68)	**13.45** (± 0.08)	5.65	2.55
Ri	48	28.86 (± 0.18)	9.73 (± 0.01)	0.7	36.72 (± 0.62)	9.58 (± 0.03)	36.60 (± 0.46)	**9.41** (± 0.02)	11.59	4.93
Sp	7	26.18 (± 0.64)	5.94 (± 0.09)	0.5	26.18 (± 0.64)	5.94 (± 0.09)	35.74 (± 0.64)	**5.75** (± 0.07)	1.64	1.64
Ti	4	4490.36 (± 139.66)	**0.78** (± 0.08)	0.5	4490.36 (± 139.66)	**0.78** (± 0.08)	5698.42 (± 147.69)	0.79 (± 0.08)	-0.09	-0.09

rante el crecimiento del conjunto (Figura 6.8(b)), puede verse que el algoritmo DW-RA presenta mayores valores de este parámetro desde las primeras iteraciones. En general, se ha observado que normalmente el algoritmo DW-RA tiene un valor medio del margen mínimo ($\overline{\rho}_{min}$) mayor que el de los algoritmos RA-se y CV RA-we.

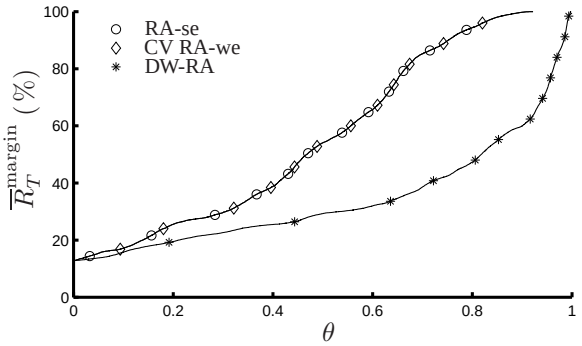

Figura 6.7: Comportamiento de $\overline{R}_T^{\text{margin}}$ $(\%)$ en función de θ para los algoritmos RA-se, CV RA-we y DW-RA en el problema *Ripley* y fijando M a 48.

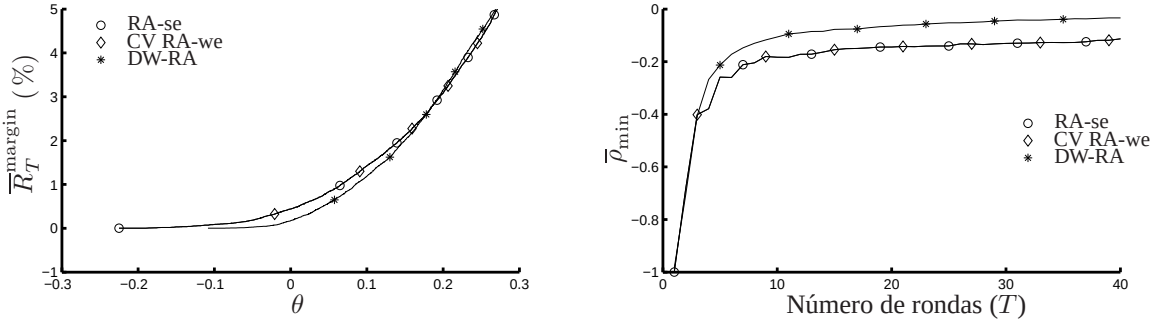

Figura 6.8: Análisis de $\overline{R}_T^{\text{margin}}(\theta)$ en el problema *Image* para $M = 11$; (a) Comportamiento de $\overline{R}_T^{\text{margin}}(\theta)$ en las cercanías de 0; (b) Evolución de $\overline{\rho}_{\min}$ con el número de rondas.

6.5.4. Aspectos adicionales

Para evaluar el comportamiento de los algoritmos RA-se, CV RA-we y DW-RA, se ha realizado un exhaustivo y extenso trabajo de simulación, analizando, además de los ya discutidos, los siguientes aspectos:

- La efectividad a la hora de seleccionar el número de neuronas ocultas (M) de los MLPs. Para analizar este aspecto, en el Cuadro 6.16 se muestra el mejor error de clasificación que cada algoritmo puede proporcionar, es decir, el que se habría obtenido si se hubiese seleccionado el valor óptimo del número de neuronas ocultas

Cuadro 6.16: Errores de clasificación de los algoritmos RA-se, CV RA-we y DW-RA cuando emplean el valor óptimo de M y además, en el caso del algoritmo CV RA-we, el valor óptimo del λ.

	RA-se		RA-we			DW-RA	
	M_0	E_{clas}	λ_0	M_0	E_{clas}	M_0	E_{clas}
Ab	6	19.14	0.1	6	19.01	**5**	**18.97**
Co	4	28.88	0	3	28.51	**2**	**28.54**
Im	9	2.29	0.3	9	2.26	15	2.13
Kw	9	11.64	0.4	9	11.64	**22**	**11.66**
Ph	54	13.65	0.1	54	13.46	**70**	**13.43**
Ri	**48**	**9.73**	1	**34**	9.30	28	9.30
Sp	5	5.78	0.6	6	5.77	5	5.63
Ti	**4**	**0.75**	0.5	4	0.75	**4**	**0.79**

(M_0) con los resultados de test (en el caso del algoritmo CV RA-we estos resultados coinciden con los de la aproximación "omnisciente"). Aunque este perverso truco no es útil para propósitos de diseño, permite saber si los algoritmos que se están evaluando realizan correctamente la selección de M.

Comparando estos resultados con los que se presentaron en el Cuadro 6.14, se puede concluir que el algoritmo DW-RA es el más robusto con respecto a la selección de M, ya que en cinco de las ocho bases de datos selecciona el valor óptimo de M, mientras que el algoritmo RA-se lo selecciona dos veces y el algoritmo CV RA-we lo hace sólo una vez, aunque sin seleccionar el valor óptimo del parámetro de mezcla (λ_0); estos casos aparecen en negrita en el Cuadro 6.16.

Además, como puede verificarse comparando los resultados correspondientes al DW-RA para distintos valores de M (compárense los resultados recogidos en los Cuadros 6.14 a 6.16), cuando este algoritmo no selecciona el valor M_0 sigue presentando resultados competitivos.

- El segundo de los aspectos que se ha evaluado es la efectividad del criterio de parada. Observando la evolución de los errores de clasificación recogidos en la Figura 6.5 y el número medio de máquinas que tiene cada conjunto (véase el Cuadro 6.14), se percibe que el criterio de parada funciona razonablemente bien para todos los algoritmos. En particular, se puede ver cómo se detiene el crecimiento de los conjuntos cuando el algoritmo ha acabado de converger, y, además, cómo se consigue evitar el sobreajuste que aparece ocasionalmente en algunos problemas.

  Estas observaciones permiten afirmar que las conclusiones alcanzadas se deben únicamente al comportamiento real de los algoritmos, y no a detener el aprendizaje en distintas fases del proceso de crecimiento.

- Por otro lado, y centrando la atención en el algoritmo DW-RA, se ha comprobado que el criterio empleado por este método para la selección dinámica de λ (i.e., seleccionar en cada paso el valor que minimiza la cota sobre el error de entrenamiento) es realmente adecuado. Para ello, se han construido una serie de conjuntos, similares a los que genera el algoritmo DW-RA, pero seleccionando en cada ronda el valor de λ que proporciona el valor del parámetro generalizado de separación que se corresponde con la mediana del conjunto de valores $\{\delta^{(j)}\}_{j=1}^{J}$ (elección muy típica en métodos de construcción de conjuntos). Los errores de clasificación obtenidos con esta aproximación no sólo son mayores que los obtenidos por el DW-RA sino que, en muchos casos, también son peores que los del algoritmo RA-se.

- En último lugar, se ha comprobado que el buen comportamiento del algoritmo DW-RA es debido al uso de diferentes tipos de énfasis y no al hecho de entrenar 11 clasificadores diferentes en cada iteración y elegir el más adecuado. Para ello, se han construido otros conjuntos RA-se, entrenando 11 clasificadores en cada iteración (todos ellos empleando $\lambda = 0.5$), añadiendo al conjunto aquél que presenta un mayor valor del parámetro generalizado de separación. Los resultados obtenidos no difieren sustancialmente de los presentados por el RA-se, lo que confirma que el éxito del DW-RA se debe realmente a la función de énfasis mixto.

Cuadro 6.17: Comités de conjuntos RA-we versus selección dinámica.

	Comités			DW-RA		T-test	Omnisciente	
	Tipo	$\overline{T}$	$\overline{E}_{clas}$	$\overline{T}$	$\overline{E}_{clas}$	$t_{Com,DW}$	$\overline{T}$	$\overline{E}_{clas}$
Ab	$\mathrm{COM_{voto}}$	153.22	19.19	37.74	**18.97**	7.48	18.70	19.01
	Selección	(± 4.38)	(± 0.02)	(± 0.28)	(± 0.02)		(± 0.01)	(± 0.00)
Co	$\mathrm{COM_{voto}}$	30.80	**25.68**	45.80	28.54	−5.99	22.58	28.51
	Selección	(± 1.96)	(± 0.44)	(± 0.62)	(± 0.18)		(± 0.07)	(± 0.02)
Im	$\mathrm{COM_{th}}$	226.48	**2.24**	31.26	2.31	−1.20	19.72	2.26
	Básico	(± 1.96)	(± 0.03)	(± 0.37)	(± 0.04)		(± 0.08)	(± 0.01)
Kw	$\mathrm{COM_{lin}}$	336.34	**11.66**	31.62	11.66	−0.33	25.56	11.64
	Bás. ó Sel.	(± 1.22)	(± 0.00)	(± 0.28)	(± 0.01)		(± 0.04)	(± 0.00)
Ph	$\mathrm{COM_{th}}$	292.22	**13.38**	38.20	13.43	−0.39	18.34	13.46
	Básico	(± 2.11)	(± 0.09)	(± 0.69)	(± 0.09)		(± 0.02)	(± 0.01)
Ri	$\mathrm{COM_{voto}}$	164.10	9.52	36.02	**9.41**	3.12	47.68	9.30
	Selección	(± 6.60)	(± 0.02)	(± 0.44)	(± 0.03)		(± 0.10)	(± 0.00)
Sp	$\mathrm{COM_{voto}}$	309.84	**5.55**	36.62	5.75	−1.96	33.04	5.77
	Básico	(± 3.13)	(± 0.07)	(± 0.55)	(± 0.07)		(± 0.17)	(± 0.01)
Ti	$\mathrm{COM_{RA-we}}$	26566.84	0.91	5698.42	**0.79**	1.03	4490.36	0.78
	Selección	(± 633.52)	(± 0.09)	(± 147.69)	(± 0.08)		(± 139.66)	(± 0.08)

6.6. CONCLUSIONES: PRESTACIONES DE LOS ALGORITMOS CON ÉNFASIS MIXTO

Para finalizar la discusión de las prestaciones de los diseños propuestos en esta Tesis, se van a comparar las tasas de error obtenidas por los comités y el algoritmo DW-RA, analizando las ventajas y limitaciones de cada uno de ellos. Además, la comparación con la aproximación "omnisciente" permitirá analizar la eficacia de ambos algoritmos a la hora de explotar las ventajas del énfasis mixto. En el Cuadro 6.17 se presentan los resultados de estos tres métodos para cada problema, y los de aplicar el T-test sobre las tasas de error de los comités y el DW-RA.

Comparando los dos algoritmos propuestos con la aproximación “omnisciente”, se comprueba que ambos algoritmos no sólo consiguen alcanzar la tasa de error presentada por el mejor conjunto RA-we, sino que, en muchos casos, la superan, lo que ilustra la eficacia de estos métodos a la hora de explotar las ventajas potenciales que la función de énfasis mixto brinda.

Si se comparan entre sí ambas propuestas, no queda claro cuál de las dos es mejor: en cuatro casos (*Im*, *Kw*, *Ph* y *Ti*) no hay diferencia estadística significativa entre sus resultados; en *Co* y *Sp* son los comités los que presentan las menores tasa de error; mientras que en los dos casos restantes casos, *Ab* y *Ri*, es el DW-RA el que presenta el menor error de clasificación. Aunque pudiera parecer que los comités de conjuntos RA-we presentan prestaciones ligeramente superiores, el algoritmo DW-RA presenta claras ventajas de orden práctico, ya que:

- El número de máquinas que deben evaluar los comités de conjuntos RA-we es, generalmente, bastante mayor que el que requieren los conjuntos DW-RA (compárense los valores de T incluidos en el Cuadro 6.17).

- Aquí se han presentado los resultados del mejor comité (entre los cuatro tipos posibles), pero, por el momento, no se dispone de ningún criterio que permita determinar qué tipo de comité conviene seleccionar en cada caso, lo que implica que, por el momento, deban explorase todas las opciones posibles.

No obstante lo anterior, no se deben olvidar las buenas prestaciones que, en general, presentan los comités lineales y con voto (sobre todo, cuando emplean el método de selección de redes), así como la mejora que se puede conseguir en los comités con “tanh”, particularmente si se emplea adecuadamente el método de clasificación acelerada.

CAPÍTULO 7

CONCLUSIONES Y LÍNEAS FUTURAS

En este último capítulo de la Tesis Doctoral se resumen sus principales aportaciones, y se sugieren algunas líneas de investigación abiertas que se consideran relevantes; varias de estas líneas ya se han iniciado, ofreciendo resultados preliminares prometedores.

7.1. CONCLUSIONES

La primera aportación de esta Tesis Doctoral es la posibilidad de emplear una función de énfasis mixto para la construcción de conjuntos de tipo Boosting. Dicho énfasis combina, mediante un parámetro de mezcla, λ, dos términos diferentes de énfasis: uno asociado al error cuadrático medio de las muestras, y otro basado en la proximidad de éstas a la frontera. Esta nueva función de énfasis permite modificar directamente el algoritmo RA, dando lugar al algoritmo RA-we ("RA with weighted emphasis"). Combinando este algoritmo con un proceso de validación cruzada (CV) para la selección del parámetro de mezcla se consiguen ventajas frente al RA, tal y como se ha comprobado experimentalmente, lo que es justificable gracias al grado de libertad adicional que aporta λ. Sin embargo, la selección de λ mediante CV no es capaz de aprovechar al máximo las posibilidades que

la nueva función de énfasis aporta, tal y como se comprueba al comparar sus prestaciones con el resultado del algoritmo RA-we cuando emplea el valor óptimo del parámetro de mezcla (aproximación "omnisciente").

Con el objetivo de realizar una selección más conveniente del parámetro de mezcla, y como segunda aportación de esta Tesis Doctoral, se han explorado dos alternativas al método de CV que permiten obtener prestaciones superiores, incluso mejores que la aproximación "omnisciente":

- La primera de ellas aprovecha la diversidad existente entre conjuntos construidos con distintos valores del parámetro de mezcla para construir comités de conjuntos RA-we. En esta línea, se han propuesto cuatro tipos diferentes de comités de redes RA-we: comités que minimizan el error cuadrático medio (con salida lineal y con activación tangente hiperbólica), comités basados en esquemas de votación generalizada y, por último, comités que emplean el criterio seguido por el RA-we para la selección de los pesos de salida.

  Junto con estas cuatro alternativas se ofrece, como aportación complementaria, la posibilidad de emplear un método de selección de redes, que permite eliminar de la combinación aquellos conjuntos RA-we con malas prestaciones, mejorando así, en la mayoría de los casos, las prestaciones del comité resultante.

- La segunda de estas alternativas, a la que se ha denominado DW-RA ("Dynamically adapted Weighted emphasis version of RA"), realiza una selección dinámica y automática del parámetro de mezcla, eligiendo en cada iteración el valor que proporciona el clasificador base con mayor parámetro de separación generalizado. Se presenta también un análisis teórico que muestra que este criterio de selección incorpora en cada iteración el clasificador base que proporciona las mejores prestaciones al conjunto en construcción, tanto en términos de velocidad de convergencia como de capacidad de generalización.

Como última aportación de esta Tesis Doctoral se encuentra un método de clasificación acelerada propuesto que reduce el tiempo de cómputo que los comités de conjuntos RA-we

requieren durante su fase operacional.

En el Capitulo 6 se mostró la validez de los comités de conjuntos RA-we y del algoritmo DW-RA, evaluando sobre una serie de bases de datos las prestaciones de ambos enfoques respecto de las proporcionadas por el RA, el CV RA-we y la aproximación "omnisciente". Las conclusiones resultantes pueden resumirse en:

- Los comités de conjuntos RA-we mejoran de manera casi sistemática las prestaciones del RA y, en la mitad de los problemas evaluados, las de la aproximación "omnisciente". Entre los diferentes tipos de comités propuestos destacan por sus prestaciones los comités lineales (minimizando el MSE) y los comités con voto, cuyo comportamiento es especialmente ventajoso cuando emplean el método de selección de redes.

- El algoritmo DW-RA también ofrece de manera sistemática tasas de error menores, o como poco similares, a las presentadas por el RA-se y el CV RA-we. Además, se han corroborado experimentalmente las ventajas que el criterio de selección dinámica de λ aporta en términos de velocidad de convergencia y capacidad de generalización frente a una selección fija como la empleada por los algoritmos RA o el CV RA-we.

- Cuando se comparan directamente los comités de conjuntos RA-we frente al DW-RA, no queda claro cuál de las dos propuestas es mejor, ya que aunque parece que los comités presentan resultados levemente mejores en un número de ocasiones ligeramente mayor, la cantidad de máquinas que deben evaluar excede considerablemente la que requieren los conjuntos DW-RA, además de que la elección de qué tipo de comité conviene utilizar no es inmediata.

- Respecto al método de clasificación acelerada, cabe destacar que consigue muy importantes reducciones del coste computacional durante la fase operacional de los comités de conjuntos RA-we y, además, es capaz de evitar el problema de sobreajuste que presentan algunos de estos comités. Esta última ventaja se ha comprobado especialmente en los comités con "tanh", donde su empleo ha reducido considerablemente las tasas de error presentadas.

Se puede concluir, por tanto, que la función de énfasis mixto propuesta proporciona ventajas indiscutibles frente al empleo de énfasis fijos, y que los métodos de combinación de conjuntos RA-we para la construcción de comités y de selección dinámica del parámetro de mezcla que emplea el algoritmo DW-RA permiten una adecuada explotación de este nuevo mecanismo atencional.

Se quiere destacar, por último, que los diseños que se están manejando tienen prestaciones sobresalientemente altas; con lo que su mejora es de muy particular relevancia, situando los diseños aquí discutidos en la punta de la investigación en clasificación máquina.

7.2. LÍNEAS DE INVESTIGACIÓN FUTURA

Del trabajo realizado se derivan tres líneas de investigación inmediatas y que prometen aportar ventajas suplementarias sobre los métodos ya propuestos:

- La primera línea de trabajo se dirige a extender los métodos de construcción de comités de conjuntos RA-we y el algoritmo DW-RA para la selección de otros parámetros de diseño. Nótese que estos métodos, tanto la construcción de comités como el DW-RA, han surgido para solventar el problema de selección del parámetro de mezcla de la función de énfasis, pero nada impide emplearlos para la selección de otros parámetros de diseño, como puede ser el número de neuronas ocultas de los MLPs, etc.

- La segunda línea consiste en la selección del parámetro suavizador del umbral, β, utilizado por el método de clasificación acelerada para comités de conjuntos RA-we. Cierto es que fijar de forma conservadora este parámetro (e.g., $\beta = 0.5$) aporta en casi todos los casos ventajas significativas en términos de coste computacional sin deteriorar las prestaciones del comité; sin embargo, una selección más fina de este parámetro permitiría un mejor aprovechamiento de las ventajas del método y, consecuentemente, puntos de trabajo más favorables en lo que se refiere al compromiso

ahorro computacional frente a prestaciones de la máquina. Para ello, se considera la posibilidad de emplear validación cruzada, que ya se ha comprobado que proporciona excelentes resultados para conjuntos RA-se [Arenas-García et al., 2007]. No obstante, debería evaluarse la eficacia de este método y otras posibles alternativas sobre comités de conjuntos RA-we.

- La última de estas líneas de investigación se orienta a encontrar un criterio que indique qué tipo de comité, entre los cuatro que se han propuesto, es el más adecuado, i.e., el que presentará mejores prestaciones en la fase operacional. Para ello podría emplearse algún parámetro (como puede ser el riesgo marginal) que evalúe, sobre el conjunto de entrenamiento, las prestaciones del comité, y a la vez sea un indicativo de su capacidad de generalización.

Por otro lado, dada la gran versatilidad de la función de énfasis propuesta en esta Tesis Doctoral, se abre una serie de posibilidades más ambiciosas para ampliar el trabajo presentado, de entre las que cabe destacar:

- Uso del énfasis mixto sobre versiones del RA adaptadas para la resolución de problemas multiclase y de regresión, así como versiones regularizadas. Estos métodos se diferencian del RA tanto en la función de coste empleada para la selección de los pesos de salida como en la función de énfasis utilizada; sin embargo, estas modificaciones son, en la mayoría de los casos, mínimas, y resulta factible extrapolar a estos tipos de redes las ideas expuestas en esta Tesis para la gestión del énfasis mixto.

- Por otro lado, la idea de emplear énfasis mixtos puede extenderse a la construcción de otro tipos de sistemas multi-red: por ejemplo, es posible pensar en realizar una combinación de elementos distinta a la lineal constante, que significa que cada uno de dichos elementos tiene una aportación de peso fijo en cualquier región del espacio de muestras que se considere; la combinación podría ser con pesos dependientes de la posición de la muestra a clasificar (como se hace en la forma tradicional de las Mezclas de Expertos), intentado optimizar simultáneamente dichos pesos y el énfasis.

Naturalmente, estas líneas, de mayor amplitud y relevancia que las inmediatamente resultantes de la Tesis, requerirán estudios y valoraciones de magnitud similar a las aquí aportadas, aún pudiendo aprovechar varios de los resultados y propuestas presentados.

APÉNDICE A

ANÁLISIS DEL RA-WE

A.1. SELECCIÓN DE LOS PESOS DE SALIDA

Para obtener los valores de los pesos de salida asociados a cada uno de los clasificadores base en el caso del algoritmo RA-we, se va a emplear el procedimiento utilizado por el algoritmo RA; de este modo, el algoritmo RA-we aparecerá como una generalización del RA original, que incluirá el caso particular $\lambda = 0.5$. Por ello, en cada ronda se seleccionará el valor de α_t que minimice la cota del error de entrenamiento dada en (2.8).

Para simplificar el desarrollo matemático en este anexo, se denotará el producto $o_t(\mathbf{x}^{(l)})d^{(l)}$ mediante $u_t(\mathbf{x}^{(l)})$, de modo que la cota a minimizar quede expresada como

$$B_t \leq \frac{1}{L}\sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] \left\{ \frac{1+u_t(\mathbf{x}^{(l)})}{2}\exp(-\alpha_t) + \frac{1-u_t(\mathbf{x}^{(l)})}{2}\exp(\alpha_t) \right\} \quad \text{(A.1)}$$

Para calcular el valor de α_t que minimiza el segundo término de (A.1), se deriva dicha expresión respecto de α_t y se iguala a 0 el resultado de la derivación:

$$\frac{1}{L}\sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] \left\{ -\frac{1+u_t(\mathbf{x}^{(l)})}{2}\exp(-\alpha_t) + \frac{1-u_t(\mathbf{x}^{(l)})}{2}\exp(\alpha_t) \right\} = 0 \quad \text{(A.2)}$$

Operando para despejar el término exponencial dependiente de α_t, se obtiene

$$\exp(2\alpha_t) = \frac{\sum_{l=1}^{L}[1 + u_t(\mathbf{x}^{(l)})] \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]}{\sum_{l=1}^{L}[1 - u_t(\mathbf{x}^{(l)})] \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]} \tag{A.3}$$

Si, además, se considera la expresión de B_t (véase la Ecuación (2.6)) en el instante $t-1$,

$$B_{t-1} = \frac{1}{L} \sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] \tag{A.4}$$

y definiendo el parámetro generalizado de separación del clasificador base como

$$\delta_t = \frac{1}{LB_{t-1}} \sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]u_t(\mathbf{x}^{(l)}) \tag{A.5}$$

se puede reescribir el cociente en (A.3), llegando a

$$\exp(2\alpha_t) = \frac{LB_{t-1} + LB_{t-1}\delta_t}{LB_{t-1} - LB_{t-1}\delta_t} = \frac{1 + \delta_t}{1 - \delta_t} \tag{A.6}$$

Finalmente, tomando logaritmos en la expresión anterior, se obtiene una expresión analítica para el cálculo de α_t

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 + \delta_t}{1 - \delta_t} \right) \tag{A.7}$$

Esta expresión es equivalente a la utilizada por el algoritmo RA, ya que γ_t puede formularse en los mismos términos que δ_t; para ello, simplemente considérese la expresión del énfasis del RA, $D_t(\mathbf{x}^{(l)})$, en función de la salida parcial del conjunto en la ronda t-ésima (véase la Ecuación (3.3)) y combínese con la expresión de B_{t-1} (Ecuación (A.4)), obteniendo así la siguiente igualdad

$$D_t(\mathbf{x}^{(l)}) = \frac{\exp(-f_{t-1}(\mathbf{x}^{(l)})d^{(l)})}{\sum_{l=1}^{L} \exp(-f_{t-1}(\mathbf{x}^{(l)})d^{(l)})} = \frac{1}{LB_{t-1}} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] \tag{A.8}$$

que permite expresar γ_t en los mismos términos de δ_t,

$$\gamma_t = \sum_{l=1}^{L} D_t(\mathbf{x}^{(l)})u_t(\mathbf{x}^{(l)}) = \frac{1}{LB_{t-1}} \sum_{l=1}^{L} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]u_t(\mathbf{x}^{(l)}) = \delta_t \tag{A.9}$$

A.2. CONVERGENCIA DEL ERROR DE ENTRENAMIENTO

Para mostrar la convergencia del error del entrenamiento del algoritmo RA-we, se va a seguir el mismo planteamiento empleado para el RA en [Schapire y Singer, 1999]. En primer lugar, considérese la relación (A.1) y opérese el término entre llaves, obteniendo

$$B_t \leq \frac{1}{2L}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]\left\{[\exp(-\alpha_t)+\exp(\alpha_t)]+u_t(\mathbf{x}^{(l)})\,[\exp(-\alpha_t)-\exp(\alpha_t)]\right\} \tag{A.10}$$

A continuación, utilícese la expresión para α_t dada en (A.7) para operar los términos exponenciales, obteniendo así

$$\exp(-\alpha_t)+\exp(\alpha_t)=\sqrt{\frac{1-\delta_t}{1+\delta_t}}+\sqrt{\frac{1+\delta_t}{1-\delta_t}}=\frac{1-\delta_t+1+\delta_t}{\sqrt{1-\delta_t^2}}=\frac{2}{\sqrt{1-\delta_t^2}} \tag{A.11}$$

$$\exp(-\alpha_t)-\exp(\alpha_t)=\sqrt{\frac{1-\delta_t}{1+\delta_t}}-\sqrt{\frac{1+\delta_t}{1-\delta_t}}=\frac{1-\delta_t-(1+\delta_t)}{\sqrt{1-\delta_t^2}}=\frac{-2\delta_t}{\sqrt{1-\delta_t^2}} \tag{A.12}$$

Sustituyendo estos resultados en (A.10) y operando, se llega a

$$\begin{aligned}
B_t &\leq \frac{1}{2L}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]\left\{\frac{2}{\sqrt{1-\delta_t^2}}+u_t(\mathbf{x}^{(l)})\frac{-2\delta_t}{\sqrt{1-\delta_t^2}}\right\}= \\
&= \frac{1}{\sqrt{1-\delta_t^2}}\frac{1}{L}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]\left\{1-u_t(\mathbf{x}^{(l)})\delta_t\right\}= \\
&= \frac{1}{\sqrt{1-\delta_t^2}}\left\{\frac{1}{L}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]-\delta_t\frac{1}{L}\sum_{l=1}^{L}\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]u_t(\mathbf{x}^{(l)})\right\}
\end{aligned} \tag{A.13}$$

Esta cota puede ser reescrita de manera más compacta recordando las expresiones de B_{t-1} y δ_t, Ecuaciones (A.4) y (A.5) respectivamente, como

$$B_t \leq \frac{1}{\sqrt{1-\delta_t^2}}\left\{B_{t-1}-B_{t-1}\delta_t^2\right\}=\sqrt{1-\delta_t^2}\quad B_{t-1} \tag{A.14}$$

y aplicando recursivamente la desigualdad, se llega a

$$B_t \leq \prod_{t'=1}^{t}\sqrt{1-\delta_{t'}^2} \tag{A.15}$$

Por último, teniendo en cuenta que $1 - x \leq \exp(-x)$, con $x > 0$ y definiendo $\delta^2 = \min_{t'=1,\ldots t}\{\delta_{t'}^2\}$, la cota dada en (A.15) puede ser, a su vez, acotada superiormente del siguiente modo

$$B_t \leq \prod_{t'=1}^{t} \sqrt{1-\delta_{t'}^2} \leq \exp\left[-\frac{1}{2}\sum_{t'=1}^{t}\delta_t^2\right] \leq \exp\left[-\frac{t}{2}\delta^2\right] \tag{A.16}$$

relación que permite, recordando que B_t es una cota superior de E_t^S (Ecuación (2.6)), obtener la expresión buscada para la cota sobre el error de entrenamiento

$$E_t^S = \frac{1}{2L}\sum_{l=1}^{L} \mid \text{sign}\left[f_t(\mathbf{x}^{(l)})\right] - d^{(l)} \mid \leq B_t \leq \prod_{t'=1}^{t}\sqrt{1-\delta_{t'}^2} \leq \exp\left[-\frac{t}{2}\delta^2\right] \tag{A.17}$$

Del mismo modo que se ha hecho en la sección anterior, esta expresión se puede particularizar para el algoritmo RA. Para ello sólo hay que considerar que γ_t admite la misma expresión que δ_t y reescribir (A.17) como

$$E_t^S = \frac{1}{2L}\sum_{l=1}^{L} \mid \text{sign}\left[f_t(\mathbf{x}^{(l)})\right] - d^{(l)} \mid \leq B_t \leq \prod_{t'=1}^{t}\sqrt{1-\gamma_{t'}^2} \leq \exp\left[-\frac{t}{2}\gamma^2\right] \tag{A.18}$$

donde $\gamma^2 = \min_{t'=1,\ldots t}\{\gamma_{t'}^2\}$.

A.3. ANÁLISIS DEL RIESGO MARGINAL

En esta sección se va a establecer una cota sobre el riesgo marginal que va a permitir analizar su comportamiento. Para ello, recuérdese, que el riesgo marginal venía dado por la expresión

$$R_T^{\text{margin}}(\theta) = \frac{1}{L}\sum_{l=1}^{L} I\left\{\rho_T(\mathbf{x}^{(l)}) \leq \theta\right\} \tag{A.19}$$

y considérese la definición del margen de clasificación de cada patrón, $\rho_T(\mathbf{x}^{(l)}) = f_T(\mathbf{x}^{(l)})d^{(l)}/\sum_{t=1}^{T}\alpha_t$, para reformular la expresión anterior en función de la salida global, es decir,

$$R_T^{\text{margin}}(\theta) = \frac{1}{L}\sum_{l=1}^{L} I\left\{f_T(\mathbf{x}^{(l)})d^{(l)} \leq \theta\sum_{t=1}^{T}\alpha_t\right\} \tag{A.20}$$

a partir de la cual se puede establecer la siguiente cota superior sobre el riesgo marginal

$$R_T^{\text{margin}}(\theta) \leq \frac{1}{L}\sum_{l=1}^{L}\exp\left[-f_T(\mathbf{x}^{(l)})d^{(l)} + \theta\sum_{t=1}^{T}\alpha_t\right] \tag{A.21}$$

nótese que para $\theta = 0$ el riesgo marginal coincide con el error de entrenamiento, y esta cota con B_t. Haciendo unas ligeras modificaciones sobre esta cota, se tiene

$$R_T^{\text{margin}}(\theta) \leq \frac{1}{L}\sum_{l=1}^{L}\exp\left[-f_T(\mathbf{x}^{(l)})d^{(l)}\right]\exp\left[\theta\sum_{t=1}^{T}\alpha_t\right] \tag{A.22}$$

donde, la primera exponencial se corresponde con B_T, por lo que, según la Ecuación (3.11), se sabe que puede acotarse por

$$B_T \leq \prod_{t=1}^{T}\sqrt{1-\delta_t^2} = \prod_{t=1}^{T}(1+\delta_t)^{\frac{1}{2}}(1-\delta_t)^{\frac{1}{2}} \tag{A.23}$$

mientras que la segunda exponencial puede expresarse en función de δ_t, para ello, simplemente hay que sustituir la expresión de α_t, i.e,

$$\exp\left[\theta\sum_{t=1}^{T}\alpha_t\right] = \prod_{t=1}^{T}\exp\left[\theta\frac{1}{2}\ln\frac{1+\delta_t}{1-\delta_t}\right] = \prod_{t=1}^{T}\left(\frac{1+\delta_t}{1-\delta_t}\right)^{\frac{\theta}{2}} \tag{A.24}$$

Y, finalmente, uniendo ambas expresiones, se llega a la expresión

$$R_T^{\text{margin}}(\theta) \leq \prod_{t=1}^{T}(1+\delta_t)^{\frac{1}{2}}(1-\delta_t)^{\frac{1}{2}}\left(\frac{1+\delta_t}{1-\delta_t}\right)^{\frac{\theta}{2}} = \prod_{t=1}^{T}(1+\delta_t)^{\frac{1+\theta}{2}}(1-\delta_t)^{\frac{1-\theta}{2}} \tag{A.25}$$

Nótese, que si en vez de considerar la cota sobre B_T para el RA-we (Ecuación (3.11)) se hubiese considerado la correspondiente al algoritmo RA (Ecuación (2.12)), y en vez de emplear la expresión de α_t correspondiente al RA-we se hubiese considerado la del RA, se habría obtenido una cota equivalente particularizada para el algoritmo RA, i.e.,

$$R_T^{\text{margin}}(\theta) \leq \prod_{t=1}^{T}(1+\gamma_t)^{\frac{1}{2}}(1-\gamma_t)^{\frac{1}{2}}\left(\frac{1+\gamma_t}{1-\gamma_t}\right)^{\frac{\theta}{2}} = \prod_{t=1}^{T}(1+\gamma_t)^{\frac{1+\theta}{2}}(1-\gamma_t)^{\frac{1-\theta}{2}} \tag{A.26}$$

A.4. ANÁLISIS DE LA FUNCIÓN DE COSTE DE LOS CLASIFICADORES BASE

En esta sección se va a analizar el segundo término de la función de coste (3.18) empleada por el RA-we para el entrenamiento de los clasificadores base; es decir, el término

$$C^{\delta}_{\lambda,t} = \sum_{l=1}^{L} D_{\lambda,t}(\mathbf{x}^{(l)}) o_t(\mathbf{x}^{(l)}) d^{(l)} \tag{A.27}$$

para poder demostrar, así, la validez de las expresiones (3.19) y (3.20).

Si, en primer lugar, se multiplica y se divide (A.27) por $\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]$, se llega a

$$C^{\delta}_{\lambda,t} = \sum_{l=1}^{L} \frac{D_{\lambda,t}(\mathbf{x}^{(l)})}{\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]} \exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] o_t(\mathbf{x}^{(l)}) d^{(l)} \tag{A.28}$$

y definiendo

$$G'_{\lambda,t}(\mathbf{x}^{(l)}) = LB_{t-1} \frac{D_{\lambda,t}(\mathbf{x}^{(l)})}{\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}]} \tag{A.29}$$

se obtiene la siguiente expresión para $C^{\delta}_{\lambda,t}$:

$$C^{\delta}_{\lambda,t} = LB_{t-1} \sum_{l=1}^{L} G'_{\lambda,t}(\mathbf{x}^{(l)}) \exp\left[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}\right] o_t(\mathbf{x}^{(l)}) d^{(l)} \tag{A.30}$$

En segundo lugar, se analiza la expresión de $G'_{\lambda,t}$; para ello, se considera por un lado la expresión de $D_{\lambda,t}$ (véase la Ecuación (3.6)) y por otro se reescribe el término exponencial de la siguiente manera

$$\exp[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)}] = D_{\lambda,t}|_{\lambda=0.5} = \frac{1}{Z'_{t-1}} \exp\left\{\frac{1}{2}[f_{t-1}(\mathbf{x}^{(l)}) - d^{(l)}]^2 - \frac{1}{2} f^2_{t-1}(\mathbf{x}^{(l)})\right\} \tag{A.31}$$

de este modo se puede escribir $G'_{\lambda,t}(\mathbf{x}^{(l)})$ como

$$\begin{aligned} G'_{\lambda,t}(\mathbf{x}^{(l)}) &= LB_{t-1} \frac{Z'_{t-1}}{Z_{t-1}} \frac{\exp\left\{\lambda_t \cdot [f_{t-1}(\mathbf{x}^{(l)}) - d^{(l)}]^2 - (1-\lambda_t) \cdot f^2_{t-1}(\mathbf{x}^{(l)})\right\}}{\exp\left\{\frac{1}{2}[f_{t-1}(\mathbf{x}^{(l)}) - d^{(l)}]^2 - \frac{1}{2} f^2_{t-1}(\mathbf{x}^{(l)})\right\}} = \\ &= LB_{t-1} \frac{Z'_{t-1}}{Z_{t-1}} \exp\left\{(\lambda_t - \frac{1}{2}) \cdot [f_{t-1}(\mathbf{x}^{(l)}) - d^{(l)}]^2 + (\lambda_t - \frac{1}{2}) \cdot f^2_{t-1}(\mathbf{x}^{(l)})\right\} = \\ &= LB_{t-1} \frac{Z'_{t-1}}{Z_{t-1}} \exp\left\{(\lambda_t - \frac{1}{2}) \cdot \left\{[f_{t-1}(\mathbf{x}^{(l)}) - d^{(l)}]^2 + f^2_{t-1}(\mathbf{x}^{(l)})\right\}\right\} \end{aligned} \tag{A.32}$$

donde el segundo término del exponente puede expresarse del modo

$$\begin{aligned}[f_{t-1}(\mathbf{x}^{(l)}) - d^{(l)}]^2 + f_{t-1}^2(\mathbf{x}^{(l)}) &= f_{t-1}(\mathbf{x}^{(l)})^2 - 2f_{t-1}(\mathbf{x}^{(l)})d^{(l)} + {d^{(l)}}^2 + f_{t-1}^2(\mathbf{x}^{(l)}) \\ &= 2\left\{ f_{t-1}(\mathbf{x}^{(l)})^2 - 2f_{t-1}(\mathbf{x}^{(l)})\frac{d^{(l)}}{2} + \frac{{d^{(l)}}^2}{4} + \frac{{d^{(l)}}^2}{4} \right\} \\ &= 2\left\{ \left[f_{t-1}(\mathbf{x}^{(l)}) - \frac{d^{(l)}}{2} \right]^2 + \frac{{d^{(l)}}^2}{4} \right\} \end{aligned} \tag{A.33}$$

lo que permite, mediante la combinación de (A.32) y (A.33), llegar a

$$G'_{\lambda,t}(\mathbf{x}^{(l)}) = LB_{t-1}\frac{Z'_{t-1}}{Z_{t-1}} \exp\left\{ 2(\lambda_t - 0.5)\left[f_{t-1}(\mathbf{x}^{(l)}) - \frac{d^{(l)}}{2} \right]^2 \right\} \exp\left\{ 2(\lambda_t - 0.5)\frac{{d^{(l)}}^2}{4} \right\} \tag{A.34}$$

Y dado que ${d^{(l)}}^2 = 1, l = 1, ..., L$, el segundo factor es constante y se puede combinar con el cociente Z'_{t-1}/Z_{t-1}, obteniendo

$$G'_{\lambda,t}(\mathbf{x}^{(l)}) = LB_{t-1}\frac{1}{Z''_{t-1}} \exp\left\{ 2(\lambda_t - 0.5)\left[f_{t-1}(\mathbf{x}^{(l)}) - \frac{d^{(l)}}{2} \right]^2 \right\} \tag{A.35}$$

Por último, se redefine esta función para que sea una función de densidad de probabilidad; para ello, se introduce un término constante, $Z_{G,t}$, que permite asegurar que $\sum_{l=1}^{L} G_{\lambda,t}(\mathbf{x}^{(l)}) = 1$, i.e.,

$$G_{\lambda,t}(\mathbf{x}^{(l)}) = \frac{1}{Z_{G,t}} \exp\left\{ 2(\lambda_t - 0.5)\left[f_{t-1}(\mathbf{x}^{(l)}) - \frac{d^{(l)}}{2} \right]^2 \right\} \tag{A.36}$$

y por lo tanto, se tiene

$$G'_{\lambda,t}(\mathbf{x}^{(l)}) = LB_{t-1}\frac{Z_{G,t}}{Z''_{t-1}} G_{\lambda,t}(\mathbf{x}^{(l)}) = \tilde{Z} G_{\lambda,t}(\mathbf{x}^{(l)}) \tag{A.37}$$

siendo $\tilde{Z}$ el agrupamiento de todos los términos constantes. Combinando la Ecuación (A.37) con la expresión de $C^{\gamma}_{t,\lambda}$ dada en (A.30), se encuentra, finalmente, la expresión de $C^{\gamma}_{t,\lambda}$ que se estaba buscando:

$$C^{\delta}_{\lambda,t} = \frac{\tilde{Z}}{LB_{t-1}} \sum_{l=1}^{L} G_{\lambda,t}(\mathbf{x}^{(l)}) \exp\left[-f_{t-1}(\mathbf{x}^{(l)})d^{(l)} \right] o_t(\mathbf{x}^{(l)})d^{(l)} \tag{A.38}$$

A.4. ANÁLISIS DE LA FUNCIÓN DE COSTE DE LOS CLASIFICADORES BASE

APÉNDICE B

BASES DE DATOS

Las prestaciones de los algoritmos propuestos en esta Tesis Doctoral se han evaluado sobre 8 bases de datos correspondientes a problemas de decisión binaria, las cuales han sido seleccionadas de modo que exista suficiente diversidad en cuanto a número de muestras, atributos y complejidad (estructura de datos), razón por la que se consideran tanto problemas reales como artificiales.

Entre ellas se encuentran cinco bases de datos pertenecientes al repositorio de la Universidad de California en Irvine (UCI) [Newman et al., 1998]; concretamente, las correspondientes a los problemas:

- *Abalone*: en su forma original, este problema consiste en estimar la edad de un tipo de caracol marino ("abalone"), a partir de una serie de medidas físicas sobre su concha (longitud, diámetro, altura, peso, etc.); sin embargo, se ha transformado en un problema de clasificación binaria, tal y como se indica en [Ruiz y López de Teruel, 2001], para determinar si la edad del molusco es superior o inferior a los 10 años.

- *Contraceptive*: el objetivo de este problema es predecir, a partir de las características demográficas y socioeconómicas de un grupo de mujeres casadas, si emplean o no

algún método anticonceptivo.

- *Image*: en este problema cada dato corresponde a una serie de medidas (valor medio de intensidad para cada componente de color, contraste, saturación, etc.) extraídas de una región de 3×3 píxeles de una serie de imágenes tomadas al aire libre. El objetivo es discriminar entre dos tipos de superficie: natural (si se trata de cielo, follaje o hierba) o artificial (si contiene ladrillo, cemento, cristal o camino).

- *Spam*: se trata de determinar si un e-mail está relacionado con información comercial no solicitada ("spam") a partir de indicadores de la frecuencia de aparición de ciertas palabras o caracteres clave.

- *Tictactoe*: esta base de datos debe su nombre al juego de las 3 en raya ("Tic-tac-toe"), ya que indica el conjunto de todas las posibles configuraciones finales en el tablero. El objetivo de la misma es predecir si el jugador que comenzó la partida ha ganado, es decir, si este jugador tiene una de las 8 posibles combinaciones para crear 3 en raya.

Se ha empleado también la siguiente base de datos obtenida de [Alinat, 1993]:

- *Phoneme*: el objetivo de este problema es distinguir entre vocales nasales y no nasales, empleando para ello las amplitudes de los cinco primeros armónicos del espectro de frecuencias, normalizados por la energía total.

Por último, se han manejado dos problemas sintéticos:

- *Kwok*: la denominación de esta base de datos se debe al nombre de su creador [Kwok, 1999]. Los datos proceden de cinco gaussianas bidimensionales esféricas: dos gaussianas para generar los datos de una clase y tres para la otra. La tasa de error de Bayes sobre el conjunto de test es del $11.3\,\%$.

- *Ripley*: propuesto por B. D. Ripley en [Ripley, 1994], los datos de cada clase proceden de una distribución bimodal formada por la mezcla de dos gaussianas de igual

matriz de covarianza y probabilidad a priori. La tasa de acierto de Bayes sobre su conjunto de test es del $8\,\%$.

Para finalizar, ha de señalarse que los problemas *Contraceptive*, *Phoneme*, *Spam* y *Tictactoe* no incluyen una partición de test predefinida, por lo que a partir de su conjunto de entrenamiento (únicos datos disponibles) se han establecido aleatoriamente 10 particiones diferentes de los conjuntos de entrenamiento y test, que contienen, respectivamente, el $60\,\%$ y $40\,\%$ del total de los datos disponibles.

Bibliografía

[Alinat, 1993] Alinat, P. (1993). Periodic Progress Report 4. Technical Thomson Report TS ASM 93/S/EGS/NC/079, ROARS Project ESPRIT II - Number 5516.

[Arenas-García et al., 2007] Arenas-García, J., Gómez-Verdejo, V., Figueiras-Vidal, A. R. (2007). Fast evaluation of neural networks via confidence rating. *Neurocomputing* (in press).

[Arenas-García et al., 2003] Arenas-García, J., Figueiras-Vidal, A. R., Sharkey, A. J. C. (2003). The beneficial effects of using multi-net systems that focus on hard patterns. *Multi Classifier Systems (Proc. 4th International Workshop), LNCS 2709*, de Windeatt, T. y Rolli, F. (editores), pp. 45–54, Surrey, UK. Springer-Verlag.

[Arenas-García et al., 2005] Arenas-García, J., Gómez-Verdejo, V., Muñoz-Romero, S., Ortega-Moral, M., Figueiras-Vidal, A. R. (2005). Fast classification with neural networks via confidence rating. *Computational Intelligence and Bioinspired Systems, 8th Intl. Work-Conference on Artificial Neural Networks, IWANN 2005, LNCS 3512*, pp. 622–629, Barcelona, España.

[Auda et al., 1995] Auda, G., Kamel, M., Raafat, H. (1995). Voting schemes for cooperative neural network classifiers. *Proc. of the IEEE International Conference on Neural Networks*, vol. 3, pp. 1240–1243, Perth, Australia.

[Bahler y Navarro, 2000] Bahler, D., Navarro, L. (2000). Methods for combining heterogeneous sets of classifiers. *17th Natl. Conf. on Artificial Intelligence (AAAI), Workshop*

on New Research Problems for Machine Learning, Austin, TX. The AAAI Press, Menlo Park, CA.

[Battiti y Colla, 1995] Battiti, R., Colla, A. M. (1995). Democracy in neural nets: voting schemes for classification. *Neural Networks*, 7(4):691–708.

[Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

[Breiman, 1999a] Breiman, L. (1999). *Combining Predictors*. Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems, pp. 31–50. Springer-Verlag, London, UK.

[Breiman, 1999b] Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11(7):1493–1518.

[Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. Wadsworth international group, Belmont, CA.

[Buntine, 1994] Buntine, W. L. (1994). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210.

[Burges, 1998] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery*, 2(2):121–167.

[Cachin, 1994] Cachin, C. (1994). Pedagogical pattern selection strategies. *Neural Networks*, 7(1):175–181.

[Choi y Rockett, 2002] Choi, S. H., Rockett, P. (2002). The training of neural classifiers with condensed datasets. *IEEE Transactions on Neural Networks*, 32(2):202–206.

[Duda et al., 2001] Duda, R. O., Hart, P. E., G.Stork, D. (2001). *Pattern Classification*. John Wiley and Sons, New York, 2nd edition.

[Freund y Schapire, 1996] Freund, Y., Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proc. of the 13th International Conference on Machine Learning*, pp. 148–156, Bari, Italy.

[Freund y Schapire, 1997] Freund, Y., Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139. An extended abstract of this work appeared in the Proc. of the 2nd European Conf. on Computational Learning Theory, March, 1995.

[Freund y Schapire, 1999] Freund, Y., Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japan Society for Artificial Intelligence*, 14(5):771–780.

[Friedman et al., 2000] Friedman, J., Hastie, T., Tibshirani, R. J. (2000). Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 38(2):337–374.

[Friedman, 2001] Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

[Fukunaga, 1990] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 2nd edition.

[Gómez-Verdejo y Figueiras-Vidal, 2006] Gómez-Verdejo, V., Figueiras-Vidal, A. R. (2006). Designing neural network committees by combining boosting ensembles. *Proc. of the 14th European Symposium on Artificial Neural Networks (ESANN)*, pp. 419–424, Bruges, Belgium.

[Gómez-Verdejo et al., 2005] Gómez-Verdejo, V., Ortega-Moral, M., Arenas-García, J., Figueiras-Vidal, A. R. (2005). Boosting by weighting boundary and erroneous samples. *Proc. of the 13th European Symposium on Artificial Neural Networks (ESSAN)*, pp. 85–90, Bruges, Belgium.

[Gómez-Verdejo et al., 2006] Gómez-Verdejo, V., Ortega-Moral, M., Arenas-García, J., Figueiras-Vidal, A. R. (2006). Boosting by weighting critical and erroneous samples. *Neurocomputing*, 69:679–685.

[Hansen y Salomon, 1990] Hansen, L. K., Salomon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001.

[Hart, 1968] Hart, P. E. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3):515–516.

[Hashem, 1995] Hashem, S. (1995). Optimal linear combination of neural networks. *Neural Networks*, 10(3):792–994.

[Haykin, 1999] Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ, 2nd edition.

[Hill y Lewicki, 2005] Hill, T., Lewicki, P. (2005). *Statistics Methods and Applications*. StatSoft, Inc., Tulsa, OK.

[Jacobs, 1995] Jacobs, R. A. (1995). Methods for combining experts'probability assessments. *Neural Computation*, 7(5):867–888.

[Jacobs et al., 1991] Jacobs, R. A., M. I., J., Nowlan, S. J., E.Hinton, G. (1991). Adaptive mixture of local experts. *Neural Computation*, 3(1):79–87.

[Jordan y Jacobs, 1992] Jordan, M. I., Jacobs, R. A. (1992). Hierarchies of adaptive experts. *Advances in Neural Information Processing Systems*, Moody, J. E., Hanson, S. J., Lippman, R. (eds.), vol. 4, pp. 985–992, San Mateo, CA. Morgan Kaufmann.

[Jordan y Jacobs, 1994] Jordan, M. I., Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214.

[Krogh y Vedelsby, 1995] Krogh, A., Vedelsby, J. (1995). Neural network ensembles, cross validation and active learning. *Advances in Neural Information Processing Systems*, Tesauro, G., Touretzky, D. S., Leen, T. K. (eds.), vol. 7, pp. 231–238, Cambridge, MA. MIT Press.

[Kuncheva, 2004] Kuncheva, L. I. (2004). *Combining Pattern Classifiers*. Wiley, New York.

[Kung y Taur, 1995] Kung, S. Y., Taur, J. S. (1995). Decision-based neural networks with signal/image classifications applications. *IEEE Transactions on Neural Networks*, 6(1):170–181.

[Kwok, 1999] Kwok, J. T. (1999). Moderating the output of support vector classifiers. *IEEE Transactions on Neural Networks*, 10(5):1018–1031.

[Luenberger, 1984] Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, 2nd edition.

[Lyhiyaoui et al., 1999] Lyhiyaoui, A., Martinez-Ramón, M., Mora-Jiménez, I., Vázquez-Castro, M., Sancho-Gómez, J. L., Figueiras-Vidal, A. R. (1999). Sample selection via clustering to construct support vector like classifiers. *IEEE Transactions Neural Networks*, 10(6):1474–1481.

[Meir, 1995] Meir, R. (1995). Bias, variance and the combination of estimators; the case of linear least squares. *Advances in Neural Information Processing Systems*, Tesauro, G., Touretzky, D. S., Leen, T. K. (eds.), vol. 7, pp. 295–302, Cambridge, MA. MIT Press.

[Meir y Rätsch, 2003] Meir, R., Rätsch, G. (2003). An introduction to boosting and leveraging. *Advanced Lectures on Machine Learning, LNCS 2600*, Mendelson, S., Smola, A. (eds.), pp. 119–184. Springer.

[Michalski, 1980] Michalski, R. (1980). Pattern recognition as rule-guided inductive inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4):349–361.

[Mora-Jimenez et al., 2003] Mora-Jimenez, I., García-Marciel, R., Figueiras-Vidal, A. R. (2003). Improving kernel-based classifiers by guided dynamic sample selection. *Proc. 13th Intl. Conf. Artificial Neural Networks in Engineering*, pp. 27–32, St. Louis, M. ASME Press.

[Munro, 1992] Munro, P. W. (1992). Repeat until bored: A pattern selection strategy. *Advances in Neural Information Processing Systems*, Moody, J. E., Hanson, S. J.,Lippman,R. (eds.), vol. 4, pp. 1001–1008, San Mateo, CA. Morgan Kaufmann.

[Newman et al., 1998] Newman, D. J., Hettich, S., Blake, C. L., Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[Nilsson, 1965] Nilsson, N. J. (1965). *Learning machines*. McGraw-Hill, New York.

[Pérez-Cruz, 2002] Pérez-Cruz, F. (2002). IRWLS Matlab toolbox to solve the SVM for pattern recognition and regression estimation. http://www.tsc.uc3m.es/∼ fernando/.

[Pérez-Cruz et al., 2001] Pérez-Cruz, F., Navia-Vázquez, A., Alarcón-Diana, P. L., Artés-Rodríguez, A. (2001). SVC-Based Equalizer for Burst TDMA Transmissions. *Signal Processing*, 81(8):1681–1693.

[Rasmussen y Williams, 1996] Rasmussen, C. E., Williams, C. K. I. (1996). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

[Ratey, 2002] Ratey, J. J. (2002). *El cerebro: manual de instrucciones*. Random House Mondadori, Barcelona.

[Ripley, 1994] Ripley, B. D. (1994). Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society*, 56(3):409–456.

[Rosen, 1996] Rosen, B. (1996). Ensemble learning using decorrelated neural networks. *Connection Science*, 8(3-4):373–384.

[Rätsch et al., 2002] Rätsch, G., Mika, S., Warmuth, M. K. (2002). On the convergence of leveraging. *Advances in Neural Information Processing Systems*, Dietterich, T., Becker, S., Ghahramani., Z. (eds.), vol. 14, pp. 487–494, Cambridge, MA. MIT Press.

[Rätsch et al., 2001] Rätsch, G., Onoda, T., Müller, K. R. (2001). Soft margins for Adaboost. *Machine Learning*, 42(3):287–320.

[Rätsch et al., 2000] Rätsch, G., Schölkopf, B., Smola, A. J., Mika, S., Onoda, T., Müller, K. R. (2000). Robust ensemble learning. *Proc. of the NIPS'98 Workshop on Large Margin Classifiers: Advances in Large Margin Classifiers*, Smola, A. J., Bartlett, P. L., Schölkopf, B., Schuurmans, D. (eds.), pp. 207–219, Cambridge, MA. MIT Press.

[Rätsch y Warmuth, 2005] Rätsch, G., Warmuth, M. K. (2005). Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152.

[Ruiz y López de Teruel, 2001] Ruiz, A., López de Teruel, P. E. (2001). Nonlinear kernel-based statistical pattern analysis. *IEEE Transactions on Neural Networks*, 12(1):16–32.

[Schapire, 1990] Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.

[Schapire et al., 1998] Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiviness of voting methods. *The Annals of Statistics*, 26(5):1651–1686.

[Schapire y Singer, 1999] Schapire, R. E., Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336.

[Scholkopf y Smola, 2002] Scholkopf, B., Smola, A. J. (2002). *Learning with kernels*. MIT Press, Cambridge, MA.

[Sharkey, 1996] Sharkey, A. J. C. (1996). On combining artificial neural nets. *Connection Science. Special Issue on Combining Artificial Neural Nets: Ensemble Approaches*, 8(3-4):299–313.

[Sharkey, 1999] Sharkey, A. J. C. (1999). *Multi-Net Systems*. Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems, pp. 1–30. Springer-Verlag, London, UK.

[Sharkey y Sharkey, 1997] Sharkey, A. J. C., Sharkey, N. E. (1997). Combining diverse neural nets. *The Knowledge Engineering Review*, 12(3):231–247.

[Sharkey et al., 1996] Sharkey, A. J. C., Sharkey, N. E., Chandroth, G. O. (1996). Diverse neural net solutions to a fault diagnosis problem. *Neural Computing and Applications*, 4(4):218–227.

[Valiant, 1984] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

[Van Trees, 1968] Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory: Part I*. Wiley, New York.

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

[Vapnik y Chervonenkis, 1971] Vapnik, V., Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

[Wann et al., 1990] Wann, M., Hediger, T., Greenbaum, N. (1990). The influence of training sets on generalization in feed-forward neural networks. *Proc. International Joint Conference on Neural Networks (IJCNN)*, vol. III, pp. 137–142, San Diego, CA.

[Xu et al., 1992] Xu, L., Krzyzak, A., Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions Systems, Man, and Cybernetics*, 22(3):418–435.